Submitted to *Theoretical Economics*

SAFE IMPLEMENTATION

MALACHY JAMES GAVAN
University of Liverpool, Management School

ANTONIO PENTA
UPF, Dept. of Econ. and Business, ICREA, BSE and TSE,

Implementation theory is concerned with the existence of mechanisms in which, at each state of the world, all equilibria result in outcomes that are within a given Social Choice Correspondence (SCC). But if agents make mistakes, if their preferences or the solution concept are misspecified, or if the designer is limited in what can be used as punishments, then it may be desirable to insist that also deviations result in 'acceptable' outcomes. *Safe Implementation* adds this extra requirement to standard implementation. Our primitives therefore also include an Acceptability Correspondence, which like the SCC maps states of the world to sets of allocations. When the underlying solution concept is Nash Equilibrium, we identify necessary and sufficient conditions (namely, *Comonotonicity* and *Safe No-Veto*) that restrict the joint behavior of the SCC and of the Acceptability Correspondence, and that generalize Maskin's (1977) conditions. In relevant economic applications, these conditions can be quite permissive. But in 'rich' preference domains, Safe Implementation is impossible, regardless of the solution concept.

KEYWORDS: Comonotonicity, Mechanism Design, Implementation, Robustness, Safe Implementation

.

JEL CLASSIFICATION. C72, D82.

Malachy James Gavan: malachy.gavan@liverpool.ac.uk
Antonio Penta: antonio.penta@upf.edu

We are grateful to the Editor and the anonymous referees for their very helpful comments. We also thank Larbi Alaoui, Mehmet Barlo, Olivier Bochet, Aygun Dalkiran, Pia Ennuschat, Alexander Frug, Ritesh Jain, Michele Lombardi, Marco Mariotti and William Sandholm, as well as several seminar and conference audiences. The BSE acknowledges the financial support of the Spanish Ministry of Economy and Competitiveness, through the Severo Ochoa Programme, CEX2019-000915-S. Antonio Penta acknowledges the financial support of the European Research Council, ERC St-G #759424.

1. INTRODUCTION

Since Maskin (1977, 1999)'s seminal work, implementation theory has played a central role in developing our understanding of market mechanisms, institutions, and their foundations. The theory starts out by specifying a set of agents, a set of states that pin down agents' preferences, and a Social Choice Correspondence (SCC) that specifies, for each state, the set of allocations that the designer wishes to induce. While commonly known by the agents, the state of nature is unknown to the designer, and hence in order to choose the allocation the designer must rely on agents' reports. The main objective of the theory is to study the conditions under which it is possible to specify a mechanism in which, at every state, the allocations selected by the SCC are sustained as the result of agents' strategic interaction. The latter is suitably modeled via game theoretic solution concepts, each giving rise to different notions of implementation.[1]

In its baseline form, the theory imposes no restriction on the mechanisms that may achieve implementation, nor on the outcomes that may arise from agents' deviations.[2] In practice, though, the designer does not always have this freedom, or perhaps not independent of the kind, the circumstances, or the number of deviations. In some contexts, especially harsh punishments may not be *acceptable*, and hence certain allocations may be used to incentivize the agents in some states of the world, but not in others; also, depending on the states, the designer himself may be able to commit to certain outcomes of the mechanism, but not to others. When these considerations are present, the insights we receive from the classical literature are not applicable. We provide some examples:

(i) In a juridical context, for instance, the viable punishments and rewards in response to 'deviant' behavior are often restricted by other constraints or desiderata, such as constitutional rights, higher level legislation, culture, or social norms.

[1]For instance, Nash (Maskin, 1999) and Subgame Perfect (Moore and Repullo, 1988), or more recently Rationalizable (Bergemann et al. (2011), Kunimoto and Serrano (2019), Kunimoto et al. (2024)), Level-k (De Clippel et al., 2019), and Behavioral (De Clippel, 2014) Implementation. Maskin and Sjöström (2002) survey the early literature. Robustness with respect to misspecification of the solution concept is studied in Jain et al. (2024).

[2]Restrictions on the mechanisms have sometimes been imposed, but by and large the literature has not paid attention to a mechanism's outcomes at profiles that are not consistent with the solution concept. Some exceptions are Bochet and Tumennasan (2023a,b), Shoukry (2019), and Eliaz (2002), which we discuss in Section 6.

(ii) A competition authority wants to induce a certain market arrangement, which depends on information that is only available to the firms, but is subject to political constraints that limit its ability to use certain punishments and rewards at certain states (see Ex. 1).

(iii) The designer may also care that the outcomes of deviations are *acceptable*, or very close the first-best 'target' allocation, if he is concerned that the agents may make mistakes, that they are boundedly rational, or that their preferences are misspecified, etc.

To account for these considerations, we enrich the baseline framework by adding an *acceptability correspondence* that specifies, for each state of the world, the set of allocations that the designer wishes to ensure, if up to k agents deviate from the profiles that are consistent with the solution concept at that state. The resulting notion of *Safe Implementation* thus requires that, besides achieving implementation, also the outcomes of up to k deviations are 'acceptable'. Besides the illustrative examples above, this notion provides a flexible framework to study a variety of robustness notions, related to a mechanism's safety and resilience properties, and it may also accommodate important and understudied problems within the implementation literature, such as the case of state-dependent feasible outcomes (Postlewaite and Wettstein 1989), limited commitment on the designer's part (as in Ex. 1 below), a variety of robustness concerns, behavioral considerations, and others.

This modeling change, however, raises a number of challenges. These are due to a tension between the elicitation of the state of the world, the outcomes that need to be implemented, and the punishments that the designer can use to discipline agents' behavior, which are state-dependent themselves. Intuitively, if achieving standard (i.e., non-safe) implementation can be thought of as providing agents with the incentives to reveal the state, through a suitable scheme of punishments and rewards, with Safe Implementation the punishments that can be used are restricted by the very information they are designed to extract. Hence, not only must agents be given the incentives to induce socially desirable allocations, but also to reveal which prizes and punishments can be used to achieve this task.

This interplay becomes apparent in the necessary and sufficient conditions that we provide, respectively in Sections 3 and 4, when the underlying solution concept is Nash Equilibrium. Our necessary condition, *Comonotonicity*, entails a joint restriction on the Social Choice and on Acceptability Correspondences. For single-valued SCC (or Social Choice Functions, SCF), for instance, if Maskin Monotonicity requires that an allocation that is selected by the SCF at one state must also be selected at any other state in which it has (weakly) climbed up in all agents' rankings of the feasible alternatives, *Comonotonicity* strengthens it in two ways: first, it states that for such an allocation to be selected by the SCF at the second state, it suffices that it climbs (weakly) up in everyone's ranking *only* compared to the alternatives that are acceptable at the first state; second, it requires the acceptability correspondence (not the SCF) to satisfy a form of monotonicity akin to Maskin's. As for sufficiency, our results show that *Comonotonicity* is almost sufficient as well, since it always ensures *Safe Implementation* in combination with a generalization of Maskin's No-Veto condition that we call *Safe No-Veto*, which is often automatically satisfied.[3] Both *Comonotonicity* and *Safe No-Veto* coincide with Maskin's conditions whenever the acceptability correspondence is vacuous, in which case Safe Implementation also coincides with (non-safe) Nash Implementation; but they are stronger in general. For the necessity part of our results, this is because the safety requirement that we impose does make implementation harder to obtain, and the conditions we provide directly reflect the extent to which this is the case.[4] Consider the following example:

EXAMPLE 1—Competition Policy with Non-Credible Punishments: Three firms, $1, 2,$ and 3, are monopolists within their respective countries. While currently active only on their local markets, firms 1 and 2 could operate in any country. Firm 3 instead is a highly indebted company, who can only operate in its own country. A competition authority needs to choose between maintaining the status quo (allocation a), or changing the level of competition in the three markets by implementing alternatives b or c. In alternative b, all firms are active on all markets they can access, which they share equally with the competing firms. Alternative c is the same as the status quo, except that the regulator lets firm 3 go bankrupt, splits 3's market equally between 1 and 2, but these firms must each pay half of the debt of firm 3.

There are three possible states for the demand in market 3, which can be low (L), medium (M) or high (H). The true state is known to the firms but not to the designer. Firms' preference orderings at each state are represented in Figure 1. The competition authority would

[3] For our general results on SCC, we distinguish between a *weak* and a *strong* version of Comonotonicity. The two notions coincide for SCF. For SCC, the first notion is necessary, the second is for sufficiency.

[4] This result highlights an important difference between our approach and Eliaz's (2002), where the restrictions on the mechanism cannot be thought of as an *extra* desideratum on top of Nash implementation: implementation in the sense of Eliaz (2002) may obtain even if Nash Implementation is impossible (see Section 6).

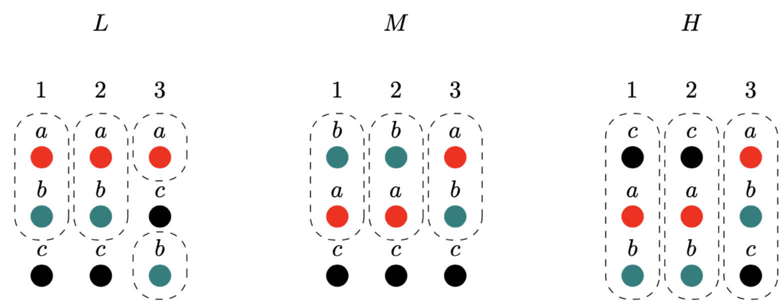

FIGURE 1.—Firms 1, 2, and 3's preference orderings over the three alternatives, at the three states, L, M, and H (e.g., firm 3's ordering at state L is $a \succ c \succ b$). The acceptability correspondence, shown in dashed lines, is such that $A(L) = A(M) = \{a, b\}$, $A(H) = \{a, b, c\}$. In this setting, the SCF such that $f(L) = a$ and $f(M) = f(H) = b$ is Nash Implementable, but not Safely so, with respect to acceptability correspondence A.

like to induce the competitive outcome, b, unless all firms prefer to maintain the status quo. Then, the SCF they wish to implement is such that $f(L) = a$ and $f(M) = f(H) = b$. Based on Maskin's results, absent safety concerns, this SCF is Nash Implementable in this setting.

But now suppose that alternative c is *not acceptable* at the states where it is at the bottom for a majority of the firms, even as the outcome of a punishment designed to implement the SCF above. This may be because it would not be desirable for the designer to let firm 3 go bankrupt, or because it would not be politically credible to commit to enforcing such an outcome, if needed, in response to someone's deviation (for instance, the three firms can be from three different European countries, and it may not be credible that the competition authority would get the political support to let country's 3 firm go burst, if needed, at a state when it is the worst outcome for the majority). That is, suppose that outcome c does not belong to the acceptability correspondence at states L and M. Then, it turns out that the SCF above cannot be Safely Implemented in this case. Thus, if the designer is subject to such political constraints, which make outcome c not credible at some states, then the insights based on the classical results are misleading.

Specifically, our results imply that in order to fulfill the Safety requirement, the designer in this case must settle for the status quo also at state H, thereby implementing a SCF that induces the competitive outcome less often. The intuition is that if b and not a has to be selected at state H (as entailed by SCF f above), in order to avoid the existence of a Nash equilibrium at H in which firms collude so as to induce the non-competitive outcome, the designer must rely on outcome c as a deterrent, since at such a state all agents prefer a over b. But if this were allowed, then c could emerge as the outcome of a deviation from an equilibrium at state L, where it is not acceptable. As a consequence, c cannot be used to discipline behavior at state H either, and hence only a SCF that chooses the same outcome at both L and H can be implemented. □

After providing the general necessary and sufficient conditions for Safe Implementation, and discussing several extensions of the main results, in Section 5 we move on to consider special cases of interest. Overall, these results show that there are important economic environments in which safety concerns can be accommodated at minimal or no cost. But Safe Implementation also has its limits: as we further show, seemingly plausible safety requirements can never be implemented, regardless of the underlying solution concept (be it Nash Equilibrium or not), when preferences are 'rich' or when the SCF is surjective on the space of feasible allocations. Thus, safety requirements are demanding in general, and there are serious limits to their implementability. Nonetheless, economically important settings exist in which they can be guaranteed under standard and generally weak conditions.

We discuss the related literature in Section 6, and conclude with Section 7, where we explain how our approach may contribute to the literature on behavioral implementation (see, e.g., Eliaz (2002), Renou and Schlag (2011), Tumennasan (2013), De Clippel (2014), De Clippel et al. (2019), Crawford (2021), etc.), both by favoring its integration with classical notions and by providing a 'detail free' way of accounting for the possibility of behavioral deviations, without necessarily ascribing to a particular theory thereof.

2. MODEL

We consider environments with complete information, with a finite set of agents, $N = \{1, ..., n\}$, and an outcome space X. Each agent i has a bounded utility $u_i : X \times \Theta \to \mathbb{R}$, where Θ is the set of states of nature, with typical element $\theta \in \Theta$, which we assume is commonly known by the agents but unknown to the designer. We let $\mathcal{E} = \langle N, \Theta, X, (u_i)_{i \in N} \rangle$ denote the environment from the viewpoint of the designer, and for any $\theta \in \Theta$, we let $\mathcal{E}(\theta) := \langle N, X, (u_i(\cdot, \theta))_{i \in N} \rangle$ denote the environment in which agents commonly know that preferences are $(u_i(\cdot, \theta))_{i \in N}$. Finally, for any $i \in N$, $\theta \in \Theta$ and $x \in X$, we let $L_i(x, \theta) := \{y \in X : u_i(y, \theta) \leq u_i(x, \theta)\}$ denote i's lower contour set of x in state θ.

A social planner aims to choose an outcome (or a set of outcomes), as a function of the state of nature. These objectives are represented by a *social choice correspondence* (SCC), $F : \Theta \to 2^X \setminus \{\emptyset\}$. The special case when $F(\theta)$ is a singleton for every θ is referred to as *social choice function* (SCF), and denoted by $f : \Theta \to X$.

A *mechanism* is a tuple $\mathcal{M} = \langle (M_i)_{i\in N}, g\rangle$, where for each $i \in N$, M_i denotes the set of messages of agent i, and $g : M \to X$ is an outcome function that assigns one allocation to each message profile, where we let $M = \times_{i\in N} M_i$ and $M_{-i} = \times_{j\neq i} M_j$. Similarly, for subsets of players $D \subset N$, we let M_D and M_{-D} denote, respectively, the set of message profiles of all agents that are inside and outside the set D. For each $\theta \in \Theta$, any mechanism $\mathcal{M} = \langle (M_i)_{i\in N}, g\rangle$ induces a complete information game $G^{\mathcal{M}}(\theta) := \langle N, (M_i, U_i^{\theta})_{i\in N}\rangle$, where M_i is the set of strategies of player i, and payoff functions are such that $U_i^{\theta}(m) = u_i(g(m), \theta)$ for all $i \in N$ and $m \in M$.

Our main focus is on the case where agents' behavior is captured by Nash equilibrium. To this end, given a mechanism $\mathcal{M}$, we let $\mathcal{C}^{\mathcal{M}}(\theta)$ denote the set of Nash equilibria of $G^{\mathcal{M}}(\theta)$. General solution concepts are discussed in Section 6.

DEFINITION 1—Implementation: *A SCC is (fully) Implementable if there exists some mechanism $\mathcal{M}$ such that $g(\mathcal{C}^{\mathcal{M}}(\theta)) = F(\theta)$ for all $\theta \in \Theta$.*[5]

Next we introduce the new primitives that are needed for *Safe Implementation*. As we discussed in the introduction, the idea is that the designer not only wishes to attain full implementation, but also ensure that the implementing mechanism has the property that, should a number of agents deviate (perhaps due to irrationality, a mistake, or because the planner's model of their preferences or of their behavior is misspecified), the mechanism still induces outcomes that the designer regards as *acceptable*. Like the 'target' allocations in the SCC, what is regarded as *acceptable* may depend on the state. This is modelled by an *acceptability correspondence*, $A : \Theta \to 2^X \setminus \{\emptyset\}$, where $A(\theta)$ denotes the set of outcomes that the social planner regards as acceptable at state θ. A natural requirement – which, in

[5]Since F is assumed to be non empty valued, the requirement $g(\mathcal{C}^{\mathcal{M}}(\theta)) = F(\theta)$ implicitly ensures existence of the solution in the implementing mechanism (i.e., $\mathcal{C}^{\mathcal{M}}(\theta)$ is non-empty for all θ). Hence, with $\mathcal{C}^{\mathcal{M}}(\theta)$ denoting the set of Nash Equilibria, this definition coincides with the standard notion of Maskin (1999).

fact, would follow immediately as a necessary condition from Def. 2 below, and which therefore we maintain throughout – is that $F(\theta) \subseteq A(\theta)$ for all $\theta \in \Theta$.

EXAMPLE 2: ***(Some Examples and Special Cases)***

1. *Minimal Safety Guarantees:* In some settings, it may be natural to require that no agent should receive their least preferred outcome, even as the result of deviations. This can be modelled letting the acceptability correspondence $A : \Theta \to 2^X \setminus \{\emptyset\}$ be *minimally safeguarding*, i.e. such that for all $\theta \in \Theta$,

$$A(\theta) = X \setminus \left\{ x \in X : \exists j \in N \quad \textit{such that} \quad x \in \operatorname*{argmin}_{x \in X} u_j(x,\theta) \setminus \operatorname*{argmax}_{x \in X} u_j(x,\theta) \right\}.$$

2. *Planner's Welfare Guarantees:* The acceptability correspondence may explicitly represent the concerns of a social planner under second best considerations. For instance, if the planner has state-dependent preferences over allocations, $W : X \times \Theta \to \mathbb{R}$, then it is natural to think about the SCC as the set of *optimal* outcomes at every state (i.e., $F(\theta) = \arg\max_{x \in X} W(x,\theta)$ for all θ), and to consider *acceptable* allocations that ensure that the planner attains at least a certain (possibly state-dependent) reservation value $\bar{w}(\theta)$. In this case, the acceptability correspondence is defined such that, for all $\theta \in \Theta$, $A(\theta) = \{x \in X : W(x,\theta) \geq \bar{w}(\theta)\}$.
3. *Perfect Safety:* Another interesting special case is when $A(\theta) = F(\theta)$ for all $\theta \in \Theta$. This is in a sense the most demanding notion of safety, in that it requires that also the deviations do not induce outcomes inconsistent with the SCC.[6]
4. *ϵ-Perfect Safety:* When X is a metric space, one reasonable restriction is that the acceptable allocations are within a given distance from the choices in the SCC or SCF. For instance, one could define $A(\theta) = \mathcal{N}_\epsilon(f(\theta))$ for all $\theta \in \Theta$, where $\mathcal{N}_\epsilon$ is an epsilon neighbourhood with respect to the metric on X. In this sense, the acceptable allocations would be close to the 'optimal' ones in the literal sense.
5. *Limited Commitment Interpretation:* The $A(\cdot)$ correspondence may also represent other constraints that the planner faces in designing the mechanism. For instance, in

[6] Earlier work of Shoukry (2019) introduced several related notions of implementation, one of which (*weak-outcome robust implementation*) coincides with Perfect Safety in our framework. For that notion he provides one impossibility result (cf. footnote 17 below). This and other related papers are discussed in Section 6.

designing punishments and rewards for the agents, the designer may be constrained in what he can commit to, and for instance mechanisms that prescribe especially harsh punishments may not be credible at certain states after a small number of deviations. Then, for each θ, $A(\theta)$ can be taken as a primitive that encompasses the set of outcomes that the planner can credibly commit to using at that state.

6. *State-Dependent Feasible Allocations:* Our framework can also be used to accommodate the case in which the very set of feasible allocations is state-dependent, and the outcomes of a mechanism are required to be feasible both on and off equilibrium. This can be accommodated within our framework by reinterpreting $A(\theta)$ as the set of allocations that are feasible at state θ.[7]

Next, let $k \in \{1, ..., n\}$ denote the *safety level* that the designer wishes to impose. That is, the maximum number of deviations from the equilibria $m^* \in \mathcal{C}^{\mathcal{M}}(\theta)$ that the designer wants to ensure they induce outcomes in $A(\theta)$, for all θ. Formally, for each k, let N_k denote the set of all subsets of N with k elements (that is, $N_k := \{C \in 2^N : |C| = k\}$), and further define a distance function $d_N(m, m') := |\{i \in N : m_i \neq m'_i\}|$ and a neighbourhood $B_k(m) := \{m' \in M : d_N(m, m') \leq k\}$, which consists of the set of message profiles m' that differ from m for at most k messages. Also, we say that $A^* : \Theta \to 2^X \setminus \{\emptyset\}$ is a *sub-correspondence* of $A : \Theta \to 2^X \setminus \{\emptyset\}$ if it is such that $A^*(\theta) \subseteq A(\theta)$ for all $\theta \in \Theta$. With this, (A, k)-Safe Implementation is defined as follows:

DEFINITION 2—(A, k) Safe Implementation: *Fix a SCC $F : \Theta \to 2^X \setminus \{\emptyset\}$, and let $A : \Theta \to 2^X \setminus \{\emptyset\}$ denote an* acceptability correspondence*, such that $F(\theta) \subseteq A(\theta)$ for all $\theta \in \Theta$. We say that F is (A, k)-Safe Implementable if there exists a mechanism $\mathcal{M} = \langle (M_i)_{i \in N}, g \rangle$ such that: (i) F is Implemented by $\mathcal{M}$ (Def. 1)), and (ii) for all $\theta \in \Theta$, $m^* \in \mathcal{C}^{\mathcal{M}}(\theta)$, and for all $m' \in B_k(m^*)$, $g(m') \in A(\theta)$.*

If, furthermore, the acceptability correspondence, A, admits no sub-correspondence A^* for which (A^*, k)-Safe Implementation is possible, then we say that A is maximally safe.

First note that, for any k, if a SCC is (A, k)-Safe Implementable, then it is $(\hat{A}, k)$-Safe Implementable for any 'more permissive' correspondence, $\hat{A} : \Theta \to 2^X \setminus \{\emptyset\}$, such that

[7]State-dependent feasibility constraints have been studied by Postlewaite and Wettstein (1989) in the context of Walrasian Implementation, but the problem has been thoroughly neglected by the subsequent literature.

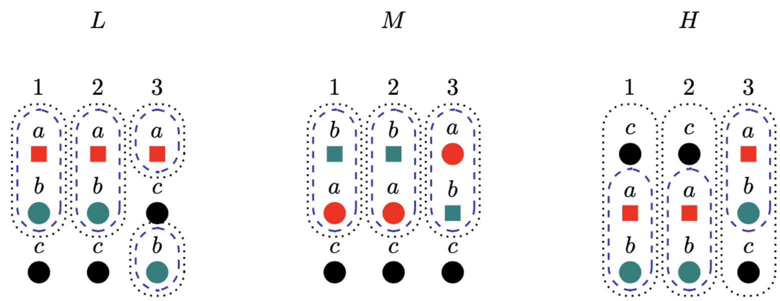

FIGURE 2.—Firms 1, 2, and 3's preference orderings over the three alternatives, at the three states, L, M, and H. For each state, the allocation chosen by SCF f^* in Ex. 3 is indicated by a square. The acceptability correspondence A from Ex.1 is shown by the dotted lines, and is not maximally safe for this SCF. Acceptability correspondence A^* in Ex. 3 is maximally safe, and is represented by the dashed lines in the figure.

$A(\theta) \subseteq \hat{A}(\theta)$ for all $\theta \in \Theta$. This observation motivates the notion of **Maximally Safe** acceptability correspondence in Def. 2: if a SCC is (A, k)-Safe Implementable, but not with respect to any sub-correspondence of A, then it means that A reflects the most demanding acceptability correspondence that can be attained.

EXAMPLE 3: Consider again the environment in Ex.1: it will follow from our results that a SCF such that $f^*(L) = f^*(H) = a$ and $f^*(M) = b$ is Safe Implementable with respect to the A correspondence in Ex.1 (see Fig.2). That acceptability correspondence, however, is not *maximally safe* for such a SCF, because it can be shown that the same SCF can also be Safe Implemented with respect to a sub-correspondence of A that rules out outcome c also at state H. Formally, $A^* : \Theta \to 2^X \setminus \{\emptyset\}$ such that $A^*(\theta) = \{a, b\}$ for all θ. $\square$

With this in mind, it should also be clear that the case $A(\theta) = F(\theta)$ for all $\theta \in \Theta$ is the most demanding one, and will be referred to as **Perfectly Safe Implementation**. We will instead use the term **Almost Perfectly Safe Implementation** to refer to the case in which, *for all* $\epsilon > 0$, Safe Implementation can be obtained with respect to an ϵ-Perfectly Safe acceptability correspondence (case 4 in Ex.2).

It is also immediate to check that if a SCC is (A, k)-Safe Implementable, then it is (A, k')-Safe Implementable for all $k' \leq k$ – that is, increasing the number of deviations the mechanism makes implementation harder – and that it always implies (baseline) Nash implementation (as we discuss in Section 6, no analogous results hold for Eliaz's (2002) concept). Also note that, when $k > 1$, Safe Implementation may accommodate the designer's concern for possibly *multi-lateral* deviations, even if the underlying solution concept is fully non-cooperative.[8]

Finally, the baseline notion in Def. 1 obtains as a special case of Def. 2 when the extra safety requirement is moot (i.e., if $A(\theta) = X$ for all $\theta \in \Theta$). In that case, Maskin (1999) showed that the following condition is necessary:

DEFINITION 3—Maskin Monotonicity: *A SCC is (Maskin) monotonic if for any θ, θ', if $x \in F(\theta)$ is such that $L_i(x, \theta) \subseteq L_i(x, \theta')$ for every $i \in N$, then $x \in F(\theta')$.*

Maskin (1999) also showed that, together with the following 'no veto condition', monotonicity is also sufficient for (baseline) Nash Implementation, whenever $n \geq 3$:

DEFINITION 4—Maskin No Veto: *A SCF satisfies the 'no veto property' if $x \in F(\theta)$ whenever $x \in X$ and $\theta \in \Theta$ are such that $\exists i \in N : \forall j \in N \backslash \{i\}, x \in \operatorname{argmax}_{y \in X} u_j(y, \theta)$.*

Obviously, Def. 4 has no bite if preferences rule out 'almost unanimity', as is the case in *economic environments*, where agents have strictly opposing interests (e.g., Mirrlees (1976), Spence (1980), Arya et al. (2000), Kartik and Tercieux (2012), etc.).

In the next two sections we provide necessary and sufficient conditions for Safe Implementation. Since Nash Implementation is a special case of Safe Implementation, the necessary conditions for Safe Implementation will have to be a generalization of Def. 3. Our sufficient conditions will also be a generalization of Maskin's, and they coincide with the necessary conditions under an 'economic condition' analogous to Kartik and Tercieux (2012)'s, or if the designer is allowed to adopt stochastic mechanisms.

3. NECESSITY

We introduce next a generalization of (Maskin) Monotonicity, which will be shown to be necessary for (A, k)-Safe Implementation:

[8] In the spirit of *renegotiation proofness*, for instance, one may want to ensure that besides implementing a SCF, the mechanism also deters joint deviations of subsets of agents. This may be achieved for instance by letting the acceptability correspondence be such that, for each $\theta \in \Theta$, no two agents prefer some $x \in A(\theta)$ over $f(\theta)$.

DEFINITION 5—Weak Comonotonicity: *A SCC,* $F:\Theta\to 2^X\setminus\{\emptyset\}$*, and an acceptability correspondence,* $A:\Theta\to 2^X\setminus\{\emptyset\}$*, are* weakly comonotonic *if the following holds:*

1. [A-Constrained Monotonicity of F] *If* $\theta,\theta'\in\Theta$ *and* $x\in F(\theta)$ *are such that* $L_i(x,\theta)\cap A(\theta)\subseteq L_i(x,\theta')\cap A(\theta)$ *for all* $i\in N$*, then* $x\in F(\theta')$.
2. [weakly F-Constrained Monotonicity of A] *If* $\theta,\theta'\in\Theta$ *are such that,* $\forall x\in F(\theta)$, $L_i(x,\theta)\cap A(\theta)\subseteq L_i(x,\theta')\cap A(\theta)$ *for all* $i\in N$*, then* $A(\theta)\subseteq A(\theta')$.

To understand this condition, first note that weak Comonotonicity implies Maskin monotonicity: If $\theta,\theta'\in\Theta$ are such that $L_i(x,\theta)\subseteq L_i(x,\theta')$, and $x\in F(\theta)$, then the condition in part 1 of Def. 5 is satisfied for any A, and hence $x\in F(\theta')$.

Second, if $A(\theta)=X$ for every θ – i.e., if the safety requirement is vacuous – then part 2 in Def. 5 holds vacuously, and part 1 coincides with (Maskin) Monotonicity. Otherwise, part 1 of Def. 5 restricts the SCC more than (Maskin) Monotonicity does. For a SCF, for instance, this condition requires that $f(\theta)=f(\theta')$ whenever $L_i(f(\theta),\theta)\cap A(\theta)\subseteq L_i(f(\theta),\theta')\cap A(\theta)$, which may be the case even if $L_i(f(\theta),\theta)\not\subseteq L_i(f(\theta),\theta')$. In the latter case, (Maskin) Monotonicity alone would leave the SCF free to set $f(\theta')\neq f(\theta)$, but weak Comonotonicity would not (see Ex. 1 in the Introduction). Thus, when the acceptability correspondence is non-trivial, weak Comonotonicity forces the SCF to be relatively more constant than Maskin's monotonicity would, and more so as the acceptability correspondence gets less permissive. More broadly, note that part 1 of Def. 5 gets less restrictive as the acceptability correspondence gets more inclusive: if A satisfies part 1 of Def. 5, and $\hat{A}$ is such that $A(\theta)\subseteq\hat{A}(\theta)$ for all $\theta\in\Theta$, then also $\hat{A}$ satisfies it.

Part 2 of Def. 5 states a monotonicity property of the acceptability correspondence, akin to Maskin's monotonicity for SCC, which imposes a lower bound on its inclusivity. Looking at the contrapositive statement, if some allocation is acceptable at state θ but not at state θ', then there must exist a 'target' allocation $x\in F(\theta)$ that, going from state θ to θ', has moved down in the ranking of the allocations within $A(\theta)$ for at least one of the agents. Note that, in this case, the bite of the condition depends on the SCC: the more inclusive the SCC, the less stringent part 2 of Def. 5. This suggests, for instance, that compared with the case of SCF, this condition leaves more freedom for the set of acceptable allocations to vary with the state when the designer aims to implement a (non single-valued) SCC.

We can now turn to our main results on necessity. As discussed in Section 2, Safe implementation becomes more restrictive as the A correspondence gets finer. Hence, as far as necessary conditions are concerned, it is natural to start with the case when the acceptability correspondence is *Maximally Safe*, which puts the most stringent constraints on safe implementation (if a SCC is (maximally) safe implementable with respect to A, then it would also be Safe-Implementable with respect to any 'coarser' acceptability correspondence, A^*, such that $A(\theta) \subseteq A^*(\theta)$ for all θ). We show next that weak Comonotonicity is necessary for *maximally* Safe Implementation:

THEOREM 1—Necessity: *A SCC, $F : \Theta \to 2^X \setminus \{\emptyset\}$, is maximally (A, k)-Safe Implementable* only if (F, A) *are weakly Comonotonic.*

To gain some intuition for this result, note that if the SCC is (A, k)-Safe Implementable and A is maximally safe, then for each $\theta \in \Theta$, $A(\theta)$ comprises *all* the outcomes that the designer can use to deter agents' deviations, and no more than those. Thus, from the viewpoint of providing agents with the right incentives within the mechanism, at any given θ, it is only agents' preferences over the set $A(\theta)$ that matter. So, if going from one state θ to another θ', one of the 'target' allocations x climbs (weakly) up in everyone's ranking *within the restricted set $A(\theta)$ of acceptable allocations* (not over all of X), and if – by the Nash implementation requirement – x must be a Nash equilibrium outcome at state θ for some mechanism, then it would also have to be a Nash equilibrium outcome at state θ'. But then x should be within the SCC at both states, otherwise Nash implementation would not obtain. This explains the necessity of part 1 of Def. 5.

To understand part 2, if going from state θ to θ' we have that *all* the allocations in $F(\theta)$ (weakly) 'climb up' in everyone's ranking within the $A(\theta)$ set, then *all* such allocations would be Nash Equilibrium outcomes at both states θ and θ', and would each be induced by some Nash equilibrium profile m^* in some mechanism. But then, in such a mechanism, the set of outcomes that are within k deviations from m^* at state θ, would also be within k-deviations from a Nash equilibrium at state θ', and thus they must also be acceptable at that state. It follows that $A(\theta')$ must contain at least all of the outcomes that are within k deviations from Nash equilibria at θ, and hence in $A(\theta)$.

As we discussed, moving to the case of non-maximally safe acceptability correspondences, Safe Implementation gets less demanding. Nonetheless, it is easy to see from the argument above that, if A is *not* maximally safe, then the first part of Def. 5 is still necessary. The second part, however, need not hold:

EXAMPLE 4: Consider again the environment in Example 3 (see Fig.2). As discussed, the SCF f^* from that example is safe implementable with respect to both correspondences A and A^*, but only the latter is *maximally safe* with respect to f^* (A cannot be, since A^* is a sub-correspondence of A). It is easy to check that, as it follows from Theorem 1, A^* satisfies both conditions in Def. 5, and hence that it is (weakly) comonotonic with respect f^*. In contrast, the A correspondence only satisfies part 1 of Def. 5 (as implied by Proposition 1), but not part 2: moving from state $\theta = H$ to $\theta' = L$, allocation $a = f^*(H)$ moves (weakly) up in everyone's ranking within the set $A(H) = \{a, b, c\}$. Yet, $A(H) \nsubseteq A(L)$. This is obviously not the case for the A^* correspondence, since $A^*(H) = A^*(L) = \{a, b\}$. □

PROPOSITION 1—Non-maximally safe implementation (necessity): $F : \Theta \to 2^X \setminus \{\emptyset\}$, *is (non-maximally)* (A, k)*-Safe Implementable* only if (F, A) *satisfy part 1 of Def. 5.*

The results above formalize a trade-off between the restrictiveness of the acceptability correspondence and the way in which the SCC correspondence varies with θ. This is easier to see considering the case of a SCF. Suppose that the designer starts with a (Maskin) Monotonic SCF. Then, among the $A^* : \Theta \to 2^X \setminus \{\emptyset\}$ correspondences that satisfy parts 1 and 2 of Def. 5, those (if they exist) that are minimal with respect to set inclusion at every state, identify the most demanding acceptability requirements that the designer can impose, if he wishes to achieve Safe Implementation. If, however, the safety desiderata are more stringent than this (i.e., if no such $\subseteq$-minimal A^* is a sub-correspondence of the acceptability correspondence that the designer wishes to impose), then Safe Implementation necessarily forces the SCF to be more constant than what is implied by (Maskin) Monotonicity (Ex.1 in the Introduction provides an instance of this).

Theorem 1 also has the following direct implication:

COROLLARY 1—Impossibility of Perfectly Safe Implementation of SCF: *For any* $k \geq 1$*, if* $f : \Theta \to X$ *and* $A : \Theta \to 2^X \setminus \{\emptyset\}$ *is such that* $A(\theta) = \{f(\theta)\}$ *for some* θ*, then* f *is*

(A,k)-Safe Implementable only if f is constant. It follows that only constant SCFs can be Perfectly Safe Implemented.[9]

This result follows directly from part 1 of Def. 5: if $A(\theta) = \{f(\theta)\}$, then $L_i(f(\theta),\theta) \cap A(\theta) = \{f(\theta)\} \subseteq L_i(f(\theta),\theta')$ for any θ', and the necessity of Comonotonicity implies that f is (A,k)-Safe Implementable only if $x = f(\theta')$ for all θ'.

Despite Corollary 1, however, in Section 5 we will show that in an important class of environments it is possible to get arbitrarily close to Perfect Safety. Specifically, under a standard single-crossing condition, Safe Implementation is possible for any (Maskin) Monotonic SCF in the *Almost Perfectly Safe* sense (i.e., for all $\epsilon > 0$, (A,k)-Safe Implementation is possible for an A-correspondence that satisfies the condition in point 4 of Ex. 2). We also stress that the negative result above holds for SCF, but as the next example shows, Perfectly Safe Implementation may be achieved if the SCC is non-single valued.

EXAMPLE 5: Let the environment be such that $\Theta = \{L,R\}$, $X = \{a,b,c\}$, $N = \{1,2,3,4\}$. Preferences are as follows: In state L, players 1 and 2 prefer a to b to c, while players 3 and 4 prefer b to c to a. In state R players 1 and 2 prefer c to b to a, while players 3 and 4 prefer a to c to b. The designer wishes to implement a SCC that selects the alternatives that are at the top of at least half of the agents (hence, $F(L) = \{a,b\}$ and $F(R) = \{a,c\}$), but ensuring *perfect safety*, in the sense that only the outcomes consistent with the SCC are deemed acceptable (that is, $A(L) = \{a,b\} = F(L)$ and $A(R) = \{a,c\} = F(R)$.) Fig. 3 summarizes as usual agents' preferences, the SCC, and the acceptability correspondence. As it will follow from Theorem 3 in the next section, such a SCC can be *perfectly safe* implemented. To see this, first notice that the intersection of player 3's lower contour set of b at state L with the acceptable allocations at that state, are not a subset of his lower contour set at state R. Hence, Comonotonicity does not require that $b \in F(R)$. Similarly, Comonotonicity does not require that $c \in F(L)$, even if $c \in F(R)$, because the relevant contour set of player 1 at state L is not a subset of that at state R. Indeed, it will be easy to verify that this environment satisfies the sufficient conditions that we provide within the next section, and hence the result will follow directly from Theorem 3. □

[9]Shoukry (2019) obtains a slightly weaker version of Corollary 1, in that $A(\theta) = \{f(\theta)\}$ is required at all states as opposed to some. The connection with Shoukry (2019) is further discssed in Section 6.

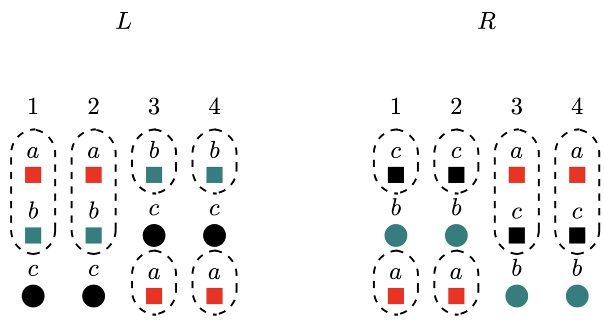

FIGURE 3.—Players 1, 2, 3 and 4's preference orderings over the three alternatives, at the two states, L and R. For each state, the allocation chosen by SCC F in Ex. 5 is indicated by a square. The acceptability correspondence A is shown by the dashed lines, and is *perfectly safe*, as it coincides with the SCC at every state.

Theorem 1 follows directly from the next result, which describes a structural property of any mechanism that safely implements the SCC. To this end, for any mechanism $\mathcal{M}$, for any $k \geq 1$, and for any $\theta \in \Theta$, let $R_k(\theta) = \bigcup_{m^* \in \mathcal{C}^{\mathcal{M}}(\theta)} B_k(m^*)$, where $\mathcal{C}^{\mathcal{M}}(\theta)$ denotes the set of Nash equilibria of $G^{\mathcal{M}}(\theta)$. That is, $R_k(\theta)$ consists of all message profiles that, given $\mathcal{M}$, are within k deviations from some Nash equilibrium at state θ. Finally, given an acceptability correspondence $A^* : \Theta \to 2^X \setminus \{\emptyset\}$ and $k \geq 1$, we say that a mechanism $\mathcal{M} = \langle (M_i)_{i \in N}, g \rangle$ is *k-surjective on A^** if, for every $\theta \in \Theta$, $g(R_k(\theta)) = A^*(\theta)$.

THEOREM 2—On the Structure of Safe Mechanisms: *Any mechanism that (A, k)-Safe Implements F must be k-surjective on some weakly Comonotonic sub-correspondence of A. If, moreover, A is* maximally *safe, then the implementing mechanism is k-surjective on A itself.*

Theorem 2 ties together the restrictions on the acceptability correspondence imposed by weak Comonotonicity, with the *safety level* parameter k. First, this result says that if a mechanism (A, k)-Safely Nash Implements F, then the A^k correspondence defined as $A^k(\theta) := g(R_k(\theta))$ for all $\theta \in \Theta$ is *weakly Comonotonic* and a sub-correspondence of A. This directly implies that A^k and F are weakly Comonotonic, and hence Theorem 1 follows from Theorem 2 when $A^k = A$, as well as the following further necessary condition for (non-maximal) Safe Implementation:

COROLLARY 2: $F : \Theta \to 2^X \setminus \{\emptyset\}$*, is (non-maximally)* (A,k)*-Safe Implementable* only if A *admits a sub-correspondence,* A^**, such that* (A^*, F) *satisfy part 2 of Def. 5.*[10]

Finally, notice that holding a mechanism $\mathcal{M}$ fixed, increasing k (weakly) enlarges the set of outcomes that are within k deviations from the Nash Equilibria at state θ, A^k. As long as the corresponding A^k defined as above is weakly Comonotonic and such that $A^k(\theta) \subseteq A(\theta)$ for all $\theta \in \Theta$, then the necessary condition for (A,k)-Safe Implementation is satisfied. But if, as k increases, the A^k correspondence is not a sub-correspondence of A, or not weakly Comonotonic, then $\mathcal{M}$ cannot (A,k)-Safe Nash implement the SCC. In that case, Safe Implementation by $\mathcal{M}$ requires either relaxing the requirement by making A more inclusive (if A^k is not a sub-correspondence of A, or if it violates part 2 of Def. 5), or to 'reduce' the dependence of the SCC on θ (if A^k violates part 1 of Def. 5). In this sense, the structural properties of any implementing 'safe' mechanism in the statement of Theorem 2 reflect a trade-off between the *safety level* parameter $k \geq 1$, the strictness of the *acceptability correspondence*, and the *responsiveness* of the SCC to the state of the world.

4. SUFFICIENCY

Our sufficiency results rely on the following stronger version of Comonotonicity:

DEFINITION 6—Strong Comonotonicity: *A SCC,* $F : \Theta \to 2^X \setminus \{\emptyset\}$*, and an acceptability correspondence,* $A : \Theta \to 2^X \setminus \{\emptyset\}$*, are* strongly comonotonic *if the following holds:*

1. [A-Constrained Monotonicity of F] *If* $\theta, \theta' \in \Theta$ *and* $x \in F(\theta)$ *are such that* $L_i(x,\theta) \cap A(\theta) \subseteq L_i(x,\theta') \cap A(\theta)$ *for all* $i \in N$*, then* $x \in F(\theta')$*.*
2. [strongly F-Constrained Monotonicity of A] *If* $\theta, \theta' \in \Theta$ *are such that* $\exists x \in F(\theta)$ *such that* $L_i(x,\theta) \cap A(\theta) \subseteq L_i(x,\theta') \cap A(\theta)$ *for all* $i \in N$*, then* $A(\theta) \subseteq A(\theta')$*.*

First, notice that the difference between *Strong* and *Weak Comonotonicity* (Def. 5) is only in the quantifier of the $x \in X$ in part 2 of the definition: in the weak version, the property $A(\theta) \subseteq A(\theta')$ is only required for states $\theta, \theta' \in \Theta$ in which $L_i(x,\theta) \cap A(\theta) \subseteq L_i(x,\theta') \cap A(\theta)$ holds for all $i \in N$ and *for all* $x \in F(\theta)$. In contrast, in Def. 6, this property

[10]Proposition 1 and Corollary 2 jointly imply that a SCC is (non-maximally) (A,k)-Safe Implementable only if A admits a weakly Comonotonic subcorrespondence. Note, however, that a non-maximally safe acceptability correspondence may still satisfy part 2 of Def. 5, i.e. with A^* in Corollary 2 such that $A^*(\theta) = A(\theta)$ for all θ.

is required to hold for all $\theta, \theta' \in \Theta$ in which $L_i(x,\theta) \cap A(\theta) \subseteq L_i(x,\theta') \cap A(\theta)$ holds for all $i \in N$ and *for some* $x \in F(\theta)$. The latter definition therefore is clearly more demanding in general, except when the SCC is single-valued (that is, when the designer wishes to implement a SCF, $f : \Theta \to X$), in which case the two notions of Comonotonicity coincide.

Strong monotonicity ensures that, when *any* allocation x that is selected at θ climbs up in the ranks for all agents when moving to θ', all acceptable allocations that are used within the mechanism to prevent deviation at θ can also be used at θ'. Our main sufficiency result will show that, under the following generalization of Maskin's No-Veto condition, *Strong Comonotonicity* is sufficient for (A,k)-Safe Implementation (in the case of SCFs, this will imply that *Comonotonicity* (either Def. 5 or 6) is both necessary and sufficient):

DEFINITION 7—Safe No-Veto: (F,A) *satisfy Safe No-Veto if* $x \in F(\theta)$ *and* $A(\theta) = X$ *whenever* $x \in X$ *and* $\theta \in \Theta$ *are such that* $\exists i, \theta' \in N \times \Theta : \forall j \in N \backslash \{i\}, x \in \text{argmax}_{y \in A(\theta')} u_j(y,\theta)$.

This property restricts both the SCC and the acceptability correspondence at states θ in which all agents but one agree that a particular allocation $x \in X$ is "best" among the set of allocations $A(\theta')$ that are acceptable at some other state θ'. At any such state, the condition requires that the SCC include such x and that all allocations be acceptable.

First note that, if the safety requirement is vacuous (i.e., if $A(\theta) = X$ for all $\theta \in \Theta$), then Def. 7 coincides with Maskin's no veto condition. In all other cases, the condition is stronger than Maskin's No-Veto for two reasons: first, because it suffices that x be at the top for 'almost everyone' only *within the set* $A(\theta') \subset X$, for some $\theta' \in \Theta$, which is implied by being at the top among *all* allocations in X, as requested by the condition for Maskin's No-Veto; second, because it entails a restriction also on the acceptability correspondence, which is required to be vacuous at least at such states θ.

THEOREM 3—Sufficiency: *If* $n \geq 3$, *and* (F,A) *are strongly Comonotonic and satisfy Safe No-Veto, then* F *is* (A,k)*-Safe Implementable for all* $k \in \mathbb{N} : 1 \leq k < \frac{n}{2}$.

Obviously, Def. 7 has no bite if preferences rule out 'almost unanimity' on any subset of allocations, as is the case in many economic settings, such as the single-crossing environments that we will consider in Section 5, or whenever the following (weaker) 'economic' restrictions hold (cf. Kartik and Tercieux (2012)):

DEFINITION 8—Economic Restrictions: *The acceptability restrictions are* economic *if, for all* $\theta, \theta' \in \Theta$ *and* $x \in X$, $\left|\{i \in N : x \in \text{argmax}_{y \in A(\theta')} u_i(y, \theta)\}\right| < n - 1$.[11]

COROLLARY 3: *If the acceptability restrictions are economic, Strong Comonotonicity of* (F, A) *is sufficient for* F *to be* (A, k)*-Safe Implementable for all* $k \in \mathbb{N} : 1 \leq k < \frac{n}{2}$.

Since Def. 5 and 6 coincide for SCFs, Theorems 1 and 3 also imply the following:

COROLLARY 4: *Let* $f : \Theta \to X$ *be such that* (f, A) *satisfy Safe No-Veto (as it is the case, for instance, under the* economic condition *in Def. 8). Then: (i)* f *is maximally* (A, k)*-Safe Nash implementable only if* (f, A) *are* Comonotonic*; (ii)* (f, A) *are* Comonotonic *only if* f *is* (A, k)*-Safe Nash implementable for all* $k \in \mathbb{N} : 1 \leq k < \frac{n}{2}$.

In the next subsections we further discuss the Safe No-Veto condition and various ways in which it can be weakened or dispensed with. The proofs of these results follow from minor adaptations of the results above, and hence we omit them. We point interested readers to the working paper version for the full proofs (Gavan and Penta (2024)).[12]

4.1. *Weakenings and Dispensability of Safe No-Veto*

Safe No-Veto holds in most standard environments, as it is unusual to have preferences where almost all agents agree. An example of this are environments that satisfy the standard single-crossing condition that we discuss in Section 5, or those that satisfy the *economic condition* in Def. 8, where Safe No-Veto can be entirely dispensed with. But even outside of these cases, under a weak 'no unanimity' condition, the requirement that $A(\theta) = X$ at those special θ can be weakened to the much more permissive condition that $A(\theta) \subseteq A(\theta')$:

DEFINITION 9—No unanimity in A: *An environment satisfies* no unanimity in A *if for all* $\theta, \theta' \in \Theta$ *and* $x \in X$, $\left|\{i \in N : x \in \text{argmax}_{y \in A(\theta')} u_i(y, \theta)\}\right| < n$.

[11] Kartik and Tercieux (2012)'s 'economic condition' obtains if $A(\theta) = X$ for all θ.

[12] We also note that, in the same sense as Maskin's No-Veto is almost necessary for Nash implementation, so is Def. 7 for Safe Implementation. The formal statement and proof can also be found in the working paper version.

DEFINITION 10—weak Safe No-Veto: (F, A) *are said to satisfy* weak Safe No-Veto *if* $x \in F(\theta)$ *and* $A(\theta) \subseteq A(\theta')$ *whenever* $x \in X$ *and* $\theta \in \Theta$ *are such that* $\exists i \in N, \theta' \in \Theta : \forall j \in N\backslash\{i\}, x \in \text{argmax}_{y \in A(\theta')} u_j(y, \theta)$.

RESULT 1—Safe Implementation under *weak* Safe No-Veto: *For any* $n \geq 3$*, if* (F, A) *are strongly Comonotonic, satisfy* no unanimity in A *and* weak Safe No-Veto*, then* F *is* (A, k)*-Safe Implementable for all* $k \in \mathbb{N} : 1 \leq k < \frac{n}{2} - 1$.

Under mild conditions on the environment, Safe No-Veto can also be dropped from the sufficient conditions via the use of a stochastic mechanism. Hence, if *stochastic mechanisms* are allowed, Strong Comonotonicity is sufficient on its own. Formally: first assume that each $u_i(\cdot, \theta)$ represents von Neumann-Morgenstern preferences, and say that a SCC is (A, k)-Safe Implementable by a stochastic mechanism if there exists $\mathcal{M} = (\langle (M_i)_{i \in I}, g \rangle$, where $g : M \rightarrow \Delta(X)$, such that (i) $\mathcal{M}$ Nash Implements the SCC and (ii) for all θ, for all Nash equilibria m^* of $G^{\mathcal{M}}(\theta)$, and for all $m \in B_k(m^*)$, $supp(g(m)) \subseteq A(\theta)$. Then, Strong Comonotonicity is sufficient under the following mild domain restriction:

DEFINITION 11: *Preferences satisfy No Total Indifference across* F *and* A *if, for all* $\theta, \theta' \in \Theta$*,* $x \in F(\theta')$ *and* $y \in A(\theta')\backslash\{x\}$*,* $\exists i \in N$ *such that* $u_i(x, \theta) \neq u_i(y, \theta)$.

RESULT 2—Safe Implementation via Stochastic Mechanisms: *Under the condition in Def. 11, for all* $n \geq 3$ *and finite* X*, if* (F, A) *are strongly Comonotonic, then* F *is* (A, k)*-Safe Implementable by a stochastic mechanism for all* $k \in \mathbb{N} : 1 \leq k < \frac{n}{2} - 1$.

For SCFs, this result immediately implies that comonotonicity (weak or strong) is both necessary and sufficient for Safe Implementation via stochastic mechanisms:

COROLLARY 5: *Let* $n \geq 3$ *and* X *be finite. Under the condition in Def. 11:* f *is maximally* (A, k)*-Safe Nash implementable by a stochastic mechanism only if* (f, A) *are* Comonotonic*; (ii)* (f, A) *are* Comonotonic *only if* f *is* (A, k)*-Safe Nash implementable by a stochastic mechanism for all* $k \in \mathbb{N} : 1 \leq k < \frac{n}{2} - 1$.

Finally, another beaten path within the literature is to consider preferences that favor 'truthfully' reporting the state and allocation (for similar ideas, see Matsushima (2008), Dutta and Sen (2012), Kartik et al. (2014), and Lombardi and Yoshihara (2020)). In this case, it can be shown that even if such *preferences for honesty* are 'weak' in the sense of being lexicographically subordinated to the outcome of the mechanism, then a mild Unanimity restriction suffices for Safe Implementation (see Gavan and Penta (2024)).

4.2. *On the Gap between Weak and Strong Comonotonicity*

Unlike Nash Implementation, where Maskin Monotonicity is both necessary and sufficient when using stochastic mechanisms under mild domain restrictions (Bochet, 2007, Benoît and Ok, 2008), a gap between necessity and sufficiency remains for Safe Implementation, since weak and strong Comonotonicity only coincide for SCFs.

In Appendix B we provide a stronger condition than Weak Comonotonicity that is necessary and almost sufficient in general environments, thereby reducing the gap between necessity and sufficiency. Similar to the 'μ Condition' in Moore and Repullo (1990), this condition relies on identifying which sub-correspondences of A are used, within an implementing mechanism, to support each of the different allocations in the SCC. Specifically, for each $x \in F(\theta)$, and for each equilibrium that induces x, we can think of the sub-correspondences of A that consist of all allocations that are within $\kappa = 1, ..., k$ deviations from such equilibrium. If, moving from θ to θ', preferences do not change *within* the sub-correspondences used to prevent *unilateral* deviations from an equilibrium that induces $x \in F(\theta)$, then x must also be implemented at θ', and hence $x \in F(\theta')$. Furthermore, the sub-correspondence of A that consists of the allocations that are reachable in k deviations from the equilibria that induce x at θ, must also be in the analogous sub-correspondence used for x at θ' (see App. B). We refer to this condition as *'Safe μ'*. But much like Moore and Repullo's μ-condition compared to Maskin Monotonicity, Safe-μ is a more complex definition to check. For this reason, we elect to provide weak and strong Comonotonicity as transparent and easy to check definitions, and leave instead this analysis for the appendix.

Turning back to stochastic mechanisms, however, it is possible provide a parallel result to Bochet (2007) and Benoît and Ok (2008). Specifically, *'Safe μ'* implies the following weaker notion of comonotonicity, which under a mild domain restriction is both necessary and sufficient for Safe Implementation via stochastic mechanisms (cf. Gavan and Penta (2024)): (A, F) satisfy *sub-comonotonicity* if there exists a correspondence G, that maps each pair (θ, x) in the graph of F to a subset of $A(\theta)$, such that if moving from state θ to

θ', an allocation $x \in F(\theta)$ 'climbs up' in the ranking of all agents within the allocations in $G(x,\theta)$, then we have that $x \in F(\theta')$ and $G(x,\theta) \subseteq G(x,\theta')$. (For a closely related condition, see Bochet and Maniquet (2010).) Note that sub-comonotonicity also boils down to Maskin's, if one takes G to be constant and equal to X.

5. APPLICATIONS AND EXTENSIONS

We now turn to two canonical applications of Nash Implementation, and include safety concerns. In the first application we explore implementation of SCFs in environments that satisfy a standard single-crossing condition. In these settings, we show that essentially any SCF can be implemented in the *Almost Perfectly Safe* sense that we discussed in p. 10. We then go on to explore the problem of allocating one unit of an indivisible good. We show that, when there is an appropriate *null allocation* that is acceptable at all states of the world, Safe Implementation of the efficient SCF is possible. Finally, we explain how our framework can accommodate arbitrary solution concepts, and we provide some negative results in environments that satisfy a strong but standard 'richness condition' on preferences.

5.1. *Environments with Private Goods and Single-Crossing Preferences*

Consider a private value settings with two private goods and single-crossing preferences.That is, for each $i \in \{1, ..., n\}$, let $X_i := \mathbb{R}^2_+$ denote the consumption space, with generic element $x_i = (x_i^1, x_i^2)$, with x_i^g denoting the quantity of good g consumed by i. The space of feasible allocations is $X \subseteq \times_{i \in N} X_i$, assumed to be compact and convex, with generic element $x = (x_i)_{i \in N}$, which is sometimes convenient to write as $x = (x_i, x_{-i})$, to separate i's own consumption bundle from the profile of consumption bundles of the others. For each agent i, there is a set of types $\Theta_i = \{\theta_i^1, ..., \theta_i^{l_i}\} \subset \mathbb{R}_+$ that pin down i's preferences over X, labelled so that $\theta_i^1 < ... < \theta_i^{l_i}$, and let $\Theta := \times_{i \in N} \Theta_i$, with typical element θ. The assumption of *private goods* is reflected in that each agent i's utility over X is constant in x_{-i}, and hence utility functions can be written as $u_i(x_i, \theta_i)$, assumed to be continuously differentiable and strictly increasing in both x_i^1 and x_i^2 for each $\theta_i \in \Theta_i$. Finally, we assume that preferences are *single-crossing* in the sense that for each i, the marginal rate of substitution between good 1 and good 2 is increasing in θ_i.

Letting $f : \Theta \to X$ denote the SCF, it seems sensible to include in the acceptability correspondence allocations that are sufficiently close to $f(\theta)$ at every $\theta \in \Theta$. (This would be natural, for instance, if the social planner chooses $f(\theta)$ to be in the argmax of a welfare functional that is continuous in x). Formally, for some $\epsilon > 0$ and neighbourhood $\mathcal{N}_\epsilon(f(\theta)) = \{(x_1, x_2) \in X : d(f(\theta), (x_1, x_2)) < \epsilon\}$, where $d(\cdot,\cdot)$ is the Euclidean distance, we assume that $\mathcal{N}_\epsilon(f(\theta)) \subseteq A(\theta)$.

LEMMA 1: *Under the maintained single-crossing condition, if $A : \Theta \to 2^X \setminus \{\emptyset\}$ is such that, for some $\epsilon > 0$, we have that $\mathcal{N}_\epsilon(f(\theta)) \subseteq A(\theta)$ for all $\theta \in \Theta$, then for* any SCF such that $f(\theta) \in \text{int}(X)$ for all $\theta \in \Theta$ then (f, A) satisfies (weak and strong) Comonotonicity.

In addition, this weak condition also suffices for Safe Implementation:

PROPOSITION 2: *Suppose that $n \geq 3$, and that the single crossing condition above is satisfied. If (f, A) is such that $f(\theta) \in \text{int}(X)$ for all $\theta \in \Theta$ and $\exists \epsilon > 0$ such that $\mathcal{N}_\epsilon(f(\theta)) \subseteq A(\theta)$ for all $\theta \in \Theta$, then f can be (A, k)-Safe Implemented for any $1 \leq k < \frac{n}{2}$.*

5.2. *Efficient Allocation of an Indivisible Good*

A social planner wants to allocate an indivisible good to one of the agents in N, or to no agent. The set of feasible outcomes therefore is $X = N \cup \{\emptyset\}$. Like Eliaz (2002), we assume that the set of states and agents' preferences are such that: (P.1) agents always prefer getting the object themselves than having it assigned to someone else; (P.2) conditional on not obtaining the object, agents always prefer it being assigned to agents with a higher utility, and prefer it not being assigned at all over being assigned to someone other than the highest utility agent; and (P.3) at any state of the world, there is always a single agent with the highest valuation.[13] Finally, we assume that the SCF and the acceptability correspondence are such that: (A.1) the SCF is efficient; (A.2) not assigning the object is always acceptable; and (A.3) whenever agent i is the designated winner, some other allocation is also acceptable.[14] Under these assumptions, the following possibility result obtains:

[13]Formally, for all i and θ: (P-1) $u_i(i,\theta) > u_i(j,\theta)$ for all $j \in N \backslash \{i\}$; (P.2) $\forall j,k \in N \backslash \{i\}$, $u_i(j,\theta) > u_i(k,\theta)$ if $u_j(j,\theta) > u_k(k,\theta)$, and $u_i(\emptyset,\theta) > u_i(j,\theta)$ if $j \notin \arg\max_{i \in N} u_i(i,\theta)$; and (P.3) $|\arg\max_{i \in N} u_i(i,\theta)| = 1$.

[14]Formally: (A.1) $f(\theta) \in \arg\max_{i \in N} u_i(i,\theta)$ for all $\theta \in \Theta$; (A.2) $\forall \theta \in \Theta$, $\{\emptyset, f(\theta)\} \subset A(\theta)$; and (A.3) For any i, whenever $f(\theta) = i$, $\exists x \neq i, \emptyset$ such that $x \in A(\theta)$.

PROPOSITION 3: *If $n \geq 3$ and preferences satisfy assumptions P.1-3, any (f, A) that satisfies assumptions A.1-3 is (A, k)-Safe Implementable for all $1 \leq k < \frac{n}{2}$.*

The assumptions on the preferences (P.1-3) are the same as in Eliaz (2002), and they are mild. Given the weakness of A.1-3, this proposition provides a rather permissive result for Safe Implementation of the efficient SCF in single-good assignment problems.

5.3. *Safe Implementation with General Solution Concepts*

Our framework can be easily extended to accommodate arbitrary solution concepts, beyond Nash equilibrium. To this end, note that for any mechanism $\mathcal{M}$, any solution concept for complete information games induces a correspondence $\mathcal{C}^{\mathcal{M}} : \Theta \to 2^M$ that assigns a (possibly empty) set of message profiles to every state of the world. So far, we took such $\mathcal{C}$ to denote the Nash Equilibrium correspondence (i.e., $\mathcal{C}^{\mathcal{M}}(\theta) := \{m^* \in M : \forall i \in N, U_i^\theta(m^*) \geq U_i^\theta(m_i, m_{-i}^*)\}$ for each θ), but both Definitions 1 and 2 extend seamlessly to any correspondence $\mathcal{C}^{\mathcal{M}} : \Theta \to 2^M$ that may be taken to model agents' strategic interaction, provided that one reinterprets notation $\mathcal{C}^{\mathcal{M}}(\theta)$ above as the set of 'solutions' (whether Nash equilibrium or not) in mechanism $\mathcal{M}$ at state θ. With this, the conceptual apparatus of Safe Implementation extends to general solution concepts: A SCC F is (A, k)-*Safe $\mathcal{C}$-Implemented* if it is $\mathcal{C}$-Implemented by a mechanism in which, at every state, any deviations of up to k agents from the profiles consistent with the solution concept $\mathcal{C}$ induce outcomes that are within the acceptability correspondence (cf. Gavan and Penta (2024)).

This general framework is useful to provide a unified view of a few related papers (which we discuss in the next Section), as well as to highlight a few methodological points regarding the agenda on behavioral implementation (which we will return to in the Conclusions). But, as we discuss next, some insights about the bite of Safety considerations may be provided independent of the solution concept, at least in environments that satisfy a richness condition analogous to the *Universal Domain* assumption in Social Choice Theory:

DEFINITION 12—Richness: *We say that Θ is* rich *if for every possible profile of strict preference orderings over X, $\succ = (\succ_i)_{i \in N}$, there exists $\theta \in \Theta$ such that $u_i(\cdot, \theta)$ represents $\succ_i$ for all $i \in N$.*

Under this condition, we provide two negative results for Safe Implementation. For the first result, take an arbitrary solution concept $\mathcal{C}$, and consider the *minimal safety guarantee* that we introduced in point 1 of Ex. 2. Under these restrictions, the social planner wishes to ensure that, in the case of deviations from the profiles admitted by the solution concept, no agent receives their least preferred outcome. This is a plausible, seemingly weak criterion for safety restrictions. Yet, under richness, we obtain the following negative result:

PROPOSITION 4: *Suppose that Θ is rich, $1 < |X| \leq n$. No SCF is (A, k)-Safe $\mathcal{C}$-Implementable for some $k \geq 1$, if A satisfies the minimal safeguarding guarantee.*

Hence, contrary to what could perhaps be surmised from the previous subsections, Safety is not a trivial restriction, regardless of the underlying solution concept.

When Nash equilibrium is taken as the underlying solution concept, as was the case in the previous sections, then this message is further reinforced by the following result: Under richness, if the SCF is onto, then the Safety requirement can only hold vacuously. Formally:

PROPOSITION 5: *Suppose that Θ is rich, and that the SCF, f, is surjective. Then, f is (A, k)-Safe(Nash) Implementable for some $k \geq 1$* only if *$A(\theta) = X$ for all θ.*

Muller and Satterthwaite (1977) showed that any SCF satisfying the above conditions must be dictatorial, and can be trivially implemented via a simple mechanism which asks the dictator for their most preferred outcome. Further to this, our result shows that all such rules require the acceptability correspondence to be vacuous. Hence, no safety considerations can be accommodated in these settings: such dictatorial rules cannot be Safe.

6. RELATED LITERATURE

The closest paper to ours is Eliaz (2002), who studies an implementation problem imposing the requirement that the mechanism's outcome is not affected by deviations of up to k agents. In that sense, the robustness desideratum in Eliaz (2002) is more demanding than ours, as it coincides with the special case of 'perfect safety', in which the acceptability correspondence coincides with the SCC (cf. point 3 in Ex. 2). Another important difference is in the solution concept: in Eliaz's (2002) k-Fault Tolerant Nash equilibrium (k-FTNE), agents reports are required to be optimal not only at the equilibrium profile, but also at all profiles in which up to k agents have deviated. Thus, the solution concept is stronger than

Nash equilibrium, and more so as k increases, with the implementation notion approaching dominant-strategy implementation as k approaches the number of opponents. This has important implications for the comparison with our approach: first, it may be that a SCC is implementable in the sense of Eliaz (2002) but not Nash Implementable – hence, unlike our notion, k-FT Implementation is not necessarily more demanding than baseline Nash Implementation; second, it may be that FT implementation is possible for some k, but not for some smaller k' – hence, unlike our notion, the implementation notion in Eliaz (2002) does not necessarily become more demanding as k increases.

In contrast, even if one replaces Nash Equilibrium in Def. 1 and 2 with a general solution concept $\mathcal{C}^{\mathcal{M}} : \Theta \to M$ (see Section 5.3), Safe Implementation always gets more demanding as k increases.[15] Fault Tolerant Implementation (FTI) fails this monotonicity because, letting $\mathcal{C}_k^{\mathcal{M}}(\theta)$ denote the set of k-FTNE at state θ, it may be that $\emptyset \neq \mathcal{C}_k^{\mathcal{M}}(\theta) \subset \mathcal{C}_{k'}^{\mathcal{M}}(\theta) \neq \emptyset$ for some $k' < k$. Thus, although k-FTNE is monotonic with respect to k (that is, all k-FTNE are also $(k-1)$-FTNE), the resulting notion of implementation is not, since the finer solution concept may make it easier to avoid the 'bad' equilibria. Hence, k-FTI does not imply $(k-1)$-FTI.[16] For the same reason, k-FTI may be more permissive than (baseline) Nash Implementation. With this, one may still ask whether (A, k)-Safe (Nash) Implementation collapses to k-FTI in the event that $A(\theta) = F(\theta)$ for all θ. This is not the case. First, contrary to k-FTI, (A, k)-Safe (Nash) Implementation is not possible for non-constant SCFs (Corollary 1). Thus, k-FTI may be more permissive than our concept, even though the two solution concepts are nested under perfect safety (i.e., when $A = F$, all k-FTNE are also (A, k)-Safe Nash equilibria). Also, for any $A : \Theta \to 2^X \setminus \{\emptyset\}$, it is not possible to have a non-constant SCF be double-implemented in k-FTI and (A, k)-Safe Nash. Finally, it can also be shown that (A, k)-Safe (Nash) Implementation may be possible when k-FTI is not (Gavan and Penta, 2024). Hence, despite the similarity in their motivation, the two implementation concepts are distinct: (i) they are not nested; (ii) unlike k-FTI, (A, k)-Safe

[15] More precisely: if the solution concept $\mathcal{C}$ does not vary with k, for any acceptability correspondence $A : \Theta \to 2^X \setminus \{\emptyset\}$, a SCC is (A, k)-Safe $\mathcal{C}$-Implementable only if it is (A, k')-Safe $\mathcal{C}$-Implementable for all $k' \leq k$.

[16] The non-monotonicity of implementation with respect to nestedness of the solution concepts is well known. It provides one of the main motivations for the notion of *strategically robust implementation* in Jain et al. (2024).

Implementation is monotonic in k; (iii) unlike k-FTI, (A, k)-Safe Implementation implies Nash Implementation. Appendix C provides examples to illustrate these points.

Eliaz (2002) also inspired Shoukry (2019), which maintains Nash equilibrium as we do, but like Eliaz (2002) only considers 'perfect safety'. As noted, this implies that the SCF is constant (cf. Corollary 1). Possibility results for non-constant SCFs are recovered allowing for transfers and a preference for the truth.[17] In contrast, here we follow the standard approach of full implementation, with standard preferences and study SCC that select subsets of the whole space of outcomes.[18]As for the safety requirement, our framework allows a wide range of acceptability correspondences, beyond the case of 'perfect safety', and we insist that *all* equilibria be safe.

Perhaps the closest to our conditions can be found in Bochet and Maniquet (2010), who study virtual implementation with support restrictions. Their *extended monotonicity* also restricts the joint behavior of two correspondences, the SCC and the (state dependent) support, in a very similar way to the sub-comonotonicity we discussed in Section 4.2. Jackson and Palfrey (2001) instead study *voluntary implementation*, with state-contingent participation constraints that can be seen as a special case of our acceptability correspondence.

Another related paper is Hayashi and Lombardi (2019), which studies Nash implementation in a two-sector economy, in which the social planner can only design the mechanism for one sector, taking the other mechanism as given. With this restriction, the possibility of preference interdependence between the two goods leads to a constraint on the planner's ability that is akin to our acceptability correspondence, because only certain allocations within the fixed sector can be achieved by deviations from a candidate equilibrium.

Postlewaite and Wettstein (1989) and Hong (1995) study continuous implementation in a Walrasian economy. They show that the implementing mechanism can be designed so that

---

[17]SCCs are also studied in Shoukry (2019), but relying on an even stronger restriction than 'perfect safety', which demands that the outcome does not change if up to k agents deviate, not just that it stays within the SCC. The concept of *weak outcome robust* implementation instead coincides with perfect safety in our framework. For this notion, he provides an impossibility result under strict unanimity and rich preferences.

[18]That is, we do not leave dimensions of the outcome space, such as transfers, outside of the SCC's codomain. Shoukry (2014) studies a distinct special case of our A-correspondence, where some agents cannot obtain alternatives that are too low in their rankings, which yields an impossibility under rich preferences. Positive results are obtained by weakening the implementation requirement so as to effectively allow some equilibria to not be safe.

the outcome function is continuous, and hence such that small deviations from the equilibria lead to small changes in the allocation, which can also be seen as a special instance of our acceptability correspondence. More broadly, also the literature on feasible implementation (Postlewaite and Wettstein, 1989, Hong, 1995, 1998) is related to our approach: as the allocations that occur upon deviations must be feasible at a given state, and feasibility constraints are state-dependent in this literature, the notion of implementation indirectly restricts the allocations that can be used upon deviations, much like Safe Implementation.

A distinct strand of literature includes concerns for robustness via changes to the solution concept. For instance, Renou and Schlag (2011) study an implementation problem where agents are unsure about the rationality of others, using a solution concept based on ϵ-minmax regret. Similarly, Tumennasan (2013) studies implementation under quantile response equilibrium, letting the logit parameter approach the perfect rationality benchmark. Barlo and Dalkıran (2021) explicitly model the possibility of preference misspecification, letting the states not pin down agents' preferences, and pursuing a notion of implementation where agents act a la Nash *for all* preferences that are consistent with each state.[19] In our paper, in contrast, we maintain Nash equilibrium and capture the possibility of mistakes (or preference misspecification) as an extra desideratum, on top of the standard notion of implementation. Bochet and Tumennasan (2023b) also maintain Nash Equilibrium, but add the extra requirement that, in a direct mechanism, not only all non-truthful profiles admit a profitable deviation (as required by baseline Nash implementation), but that deviating to truthful revelation is profitable in such instances. This notion is motivated by *resilience* considerations. A related notion can be found in De Clippel (2014), where the designer takes into account that agents may display specific deviations from rationality. For further recent approaches to behavioral implementation, see De Clippel et al. (2019), Crawford (2021), Kneeland (2022), Barlo and Dalkıran (2023), and Bochet and Tumennasan (2023a).

Finally, our results are also connected with the literature on implementation with evidence (e.g., Kartik and Tercieux (2012), Ben-Porath et al. (2019)), which also enriches

[19] In that sense, Barlo and Dalkıran (2021) can be seen as an original take on the broader idea of robust implementation, where the types that are relevant for the allocation rule pin down agents' preferences, but not their beliefs, which however matter since implementation is required to be achieved for all beliefs consistent with the designer's information (cf. in Bergemann and Morris (2005, 2009a,b), Ollár and Penta 2017, 2022, 2023).

the baseline framework with an extra feature, the ability to produce evidence. Similar to our Comonotonicity, their main conditions are also suitably adjusted versions of monotonicity. Unlike ours however, their conditions are more permissive than Maskin's (1977), effectively restricting the set of states over which monotonicity is required.

7. CONCLUSIONS

We introduce *Safe Implementation*, a notion that adds to the standard implementation requirements the restriction that deviations from the baseline solution concept induce outcomes that are *acceptable*. This is modelled by introducing, next to the Social Choice Correspondence (which represents the 'first best' objectives when agents behave in accordance with the solution concept), an Acceptability Correspondence that assigns to each state of the world the set of allocations that are considered acceptable. This framework generalizes standard notions of implementation and can accommodate a variety of questions, including robustness with respect to mistakes in play, model misspecification, behavioral considerations, state-dependent feasibility restrictions, limited commitment, etc.

Robustness concerns for mistakes in play and other behavioral considerations have been considered in the literature, mainly through changes to the solution concept (e.g., Eliaz (2002), Renou and Schlag (2011), Tumennasan (2013), De Clippel (2014), De Clippel et al. (2019), Crawford (2021), etc.) Our approach differs mainly in that we impose restrictions also on the outcomes of players' deviations, and may thus be adopted to capture concerns for misspecification of agents' behavior of any kind, as something which can be superimposed on *any* solution concept, be it 'classical' or 'behavioral' (see Section 5.3). This way, our framework can also be used to accommodate broad robustness concerns, to account for the possibility that even a behavioral model, which may have been developed in order overcome certain limitations of 'classical' notions, may of course also be misspecified. This modeling innovation therefore has the further advantage of addressing the frequent critique of behavioral models, of being *ad hoc*: in our approach, the deviations that are the object of Safety considerations are unrestricted in their nature, and hence model-free.

Decoupling these concerns from the outcomes of the solution concept, however, raises some challenges: on the one hand, like in the standard approach, the outcomes that ensue from deviations must provide the agents with the incentives to induce socially desirable outcomes, consistent with the criteria that are embedded in the underlying solution concept;

on the other hand, our concerns for safety limit precisely the designer's ability to specify such outcomes. The fact that the acceptable allocations are themselves state-dependent, like the SCC, means that not only must agents be given the incentives to induce socially desirable allocations, but also to reveal which outcomes can be used as punishments to achieve this objective. Our main results, which refer to Nash equilibrium as the underlying solution concept, precisely formalize this interplay: the necessary and sufficient conditions that we provide entail joint restrictions on the structure of the SCC and of the acceptability correspondence, and formally generalize the standard conditions for baseline Nash Implementation (Maskin, 1999). While we also offer some results for general solution concepts, that identify substantive limits to the possibility of achieving non-trivial Safety desiderata, a systematic exploration of solution concepts other than Nash equilibrium is beyond the scope of this paper, and provides an interesting direction for future research in this area.

Our framework is also general in the specification of the acceptability correspondence, which can be used to accommodate different special cases, which include: (i) "perfectly Safe implementation", which deems acceptable only the outcomes of the SCC (e.g. Eliaz (2002)); (ii) " almost perfectly Safe implementation", when only outcomes that are arbitrarily close to those in the SCC are acceptable, which provides a connection with the literature on continuous implementation (e.g., Postlewaite and Wettstein (1989), Hong (1995)); (iii) state-dependent feasibility constraints (e.g., Postlewaite and Wettstein (1989), Hong (1995, 1998)); (iv) minimal guarantees based on a variety of welfare criteria (cf. Ex. 2); (v) limited commitment in mechanism design, if the designer can only commit to carrying through, depending on the state, certain punishments but not others (cf. Ex. 1); etc. But these are only some of the possibilities that can be cast within our framework. Further exploring these or other special cases, explicitly tailored to address specific concerns in more applied settings, may provide another promising direction for future research.

Finally, as it is customary when conceptual innovations are introduced within implementation theory, we have maintained the complete information assumption and imposed no further restrictions on the mechanisms. Combining safety considerations with incomplete information, or with other restrictions on the mechanisms (e.g., Jackson (1991, 1992), Ollár and Penta (2017, 2022, 2023), etc.), is yet another direction for future research.

APPENDIX A: PROOFS

Proof of Theorem 1: Suppose that F is (A,k)-Safe Implementable. Further, suppose that it is maximally so. Therefore there is some mechanism $\mathcal{M}$ that (A,k)-Safe Implements F and is such that $A(\theta) = g(\{m \in M | d(m,m^*) \leq k, \quad m^* \in \mathcal{C}^{\mathcal{M}}(\theta)\})$.

We will show that F and A are weakly comonotonic in two steps.

Firstly, we will show that if for some $\theta, \theta' \in \Theta$, if there exists $x \in F(\theta)$ such that $L_i(x,\theta) \cap A(\theta) \subseteq L_i(x,\theta') \cap A(\theta)$ for all $i \in N$, then $x \in F(\theta')$. To do so, take m^* to be a Nash Equilibrium at θ that induces x. Hence $g(m^*) = x \in F(\theta)$. Let $\theta' \in \Theta$ be a state such that $x \notin F(\theta')$. Therefore m^* is not a Nash Equilibrium at θ' and hence $\exists i \in N$, $m_i' \in M_i$ such that $u_i(g(m_i', m_{-i}^*), \theta') > u_i(x, \theta')$. It follows that $g(m_i', m_{-i}^*) \in X \setminus L_i(x,\theta')$ and $g(m_i', m_{-i}^*) \in g(\{m \in M | d(m,m^*) \leq k, \quad m^* \in \mathcal{C}^{\mathcal{M}}(\theta)\}) = A(\theta)$. However, as m^* is a NE at θ we have that $g(m_i', m_{-i}^*) \in L_i(x,\theta) \cap A(\theta)$. Therefore it cannot be the case that $L_i(x,\theta) \cap A(\theta) \subseteq L_i(x,\theta') \cap A(\theta)$, a contradiction.

Now we show that if for some $\theta, \theta' \in \Theta$, all $x \in F(\theta)$ are such that $L_i(x,\theta) \cap A(\theta) \subseteq L_i(x,\theta') \cap A(\theta)$ for all $i \in N$, then $A(\theta) \subseteq A(\theta')$. Suppose that θ and θ' are states such that $L_i(x,\theta) \cap A(\theta) \subseteq L_i(x,\theta')$ $\forall i \in N$ for all $x \in F(\theta)$. Suppose by contradiction that $A(\theta) \nsubseteq A(\theta')$, and let m^* be a Nash Equilibrium at θ that induces $x \in F(\theta)$.

We consider two cases: (1) If m^* is a Nash Equilibrium at θ', then $B_k(m^*) \subseteq A(\theta')$ by definition. (2) If m^* is not a Nash Equilibrium at θ'. In this case, there must be some $i \in N$, who at the state θ' has a profitable deviation from m^*, i.e. $u_i(g(m_i', m_{-i}^*), \theta') > u_i(x, \theta')$. We conclude that $g(m_i', m_{-i}^*) \in X \setminus L_i(x,\theta')$. By (A,k)-Safe Implementation, and by definition we have that $A(\theta) = g(\{m \in M | d(m,m^*) \leq k, \quad m^* \in \mathcal{C}^{\mathcal{M}}(\theta)\})$, it must be that $g(m_i', m_{-i}^*) \in L_i(x,\theta) \cap A(\theta)$. A contradiction to $L_i(x,\theta) \cap A(\theta) \subseteq L_i(x,\theta')$ for all $x \in F(\theta)$.

We conclude that all m^* that are Nash Equilibria at θ and induce x, are also Nash Equilibria at θ'. Now notice that if this holds for all $y \in F(\theta)$ then all Nash Equilibria at θ are also Nash Equilibria at θ'. Given this, the outcomes induced by k agents deviating from Equilibrium at θ are also reached within k deviations of an Equilibrium at θ', and hence $A(\theta) \subseteq A(\theta')$. Thus, (F,A) must be weakly comonotonic. ■

Proof of Proposition 1: Suppose that F is (A,k)-Safe Implementable. Therefore there is some mechanism $\mathcal{M}$ that (A,k)-Safe Implements F . We will show that if for some

$\theta, \theta' \in \Theta$, if there exists $x \in F(\theta)$ such that $L_i(x,\theta) \cap A(\theta) \subseteq L_i(x,\theta') \cap A(\theta)$ for all $i \in N$, then $x \in F(\theta')$. That is, A-Constrained Monotonicity of F is satisfied. To do so, take m^* to be a Nash Equilibrium at θ that induces x. Hence $g(m^*) = x \in F(\theta)$. Let $\theta' \in \Theta$ be a state such that $x \notin F(\theta')$. Therefore m^* is not a Nash Equilibrium at θ' and hence $\exists i \in N$, $m_i' \in M_i$ such that $u_i(g(m_i', m_{-i}^*), \theta') > u_i(x, \theta')$. It follows that $g(m_i', m_{-i}^*) \in X \backslash L_i(x, \theta')$ and $g(m_i', m_{-i}^*) \in g(\{m \in M | d(m, m^*) \leq k, \quad m^* \in \mathcal{C}^{\mathcal{M}}(\theta)\}) \subseteq A(\theta)$ by definition of Safety. However, as m^* is a NE at θ we have that $g(m_i', m_{-i}^*) \in L_i(x,\theta) \cap A(\theta)$. Therefore it cannot be the case that $L_i(x,\theta) \cap A(\theta) \subseteq L_i(x,\theta') \cap A(\theta)$. ■

Proof of Theorem 2: Suppose that F is (A,k)-Safe Implementable. Therefore there is some mechanism $\mathcal{M}$ that (A,k)-Safe Implements F and is such that $g(\{m \in M | d(m, m^*) \leq k, \quad m^* \in \mathcal{C}^{\mathcal{M}}(\theta)\}) \subseteq A(\theta)$. Take A^* to be a sub-correspondence of A such that $g(\{m \in M | d(m, m^*) \leq k, \quad m^* \in \mathcal{C}^{\mathcal{M}}(\theta)\}) = A^*(\theta)$ for all states. By definition, $\mathcal{M}$ is k-surjective on A^*. Moreover, for maximal safety, we require $A^*(\theta) = A(\theta)$ for all θ, else some alternatives could be removed, contradicting maximally safe. With this, the logic of theorem 1 holds exactly, as the proof only relies on the outcomes obtainable within k deviations of the implementing mechanism. That is, one could replace $A(\theta)$ with $A^*(\theta)$ throughout. ■

Proof of Theorem 3: For each $i \in N$, let $M_i = \bigcup_{\theta' \in \Theta} A(\theta') \times \Theta \times \mathbb{N}$, with typical element $m_i = (x^i, \theta^i, n^i)$. Let $g(m)$ be as follows:

(i) If $m_i = (x, \theta, n^i)$ $\forall i \in N$ and $x \in F(\theta)$ then $g(m) = x$

(ii) If $m_i = (x, \theta, n^i)$ $\forall i \in N \backslash \{j\}$ with $x \in F(\theta)$ and $m_j = (y, \cdot, \cdot)$ then

$$g(m) = \begin{cases} y & \text{if } y \in L_j(x,\theta) \cap A(\theta) \\ x & \text{if } y \notin L_j(x,\theta) \cap A(\theta) \end{cases}$$

(iii) if $k > 1$ and $m_i = (x, \theta, \cdot)$, $x \in F(\theta)$, $\forall i \in N \backslash D$, $2 \leq |D| \leq k$ s.t. $\forall j \in D$ $m_j \neq (x, \theta, \cdot)$

$$g(m) = \begin{cases} x^{i^*} & \text{if } D^*(\theta, D) \neq \emptyset \\ x & \text{if } D^*(\theta, D) = \emptyset \end{cases}$$

where $D^*(\theta, D) = \{j \in D | x^j \in A(\theta)\}$, $i^* = \min\{i \in D^*(\theta, D) | n^i \geq n^j \quad j \in D^*(\theta, D)\}$

(iv) Otherwise, let $g(m) = x^{i^*}$ where $i^* = \min\{i \in N | n^i \geq n^j \quad \forall j \in N\}$

From here we can complete the proof in three steps: showing that all $x \in F(\theta)$ are induced by a Nash Equilibrium at θ, showing that there is no $y \notin F(\theta)$ such that y is induced by an Equilibrium at θ, and finally showing that the mechanism is indeed (A, k)-Safe.

Step 1. First we show that all $x \in F(\theta)$ are induced by Nash Equilibria at θ. Consider m^* s.t. $m_i^* = (x, \theta, \cdot), \quad \forall i \in N$ where $x \in F(\theta)$ at the state θ. To be a Nash Equilibrium we need to rule out the possibility that $\exists j \in N, m_j' \in M_j$ s.t. $u_j(g(m_{-j}^*, m_j'), \theta) > u_j(g(m^*), \theta)$. However, $g(m_{-j}^*, m_j') = y$ must be s.t. $y \in L_j(x, \theta)$ by rule (ii), therefore it is not possible that $u_j(y, \theta) > u_j(x, \theta)$. Hence, m^* is a Nash Equilibrium leading to $x \in F(\theta)$.

Step 2. We show there is no Nash equilibrium m^* at θ such that $g(m^*) = y \notin F(\theta)$.

Case 1. Suppose m^* is a Nash equilibrium in rule i) at state θ such that $g(m^*) = y \notin F(\theta)$. It must be that $m_i^* = (y, \theta', n^i)$ for all $i \in N$ and, necessarily as $y \notin F(\theta)$, that $\theta' \neq \theta$. Given this, it must be that there is no profitable deviation as m^* is a Nash equilibrium. As deviations may only lead to rule (ii), it must be that for all $i \in N$, for any $z \in L_i(y, \theta') \cap A(\theta')$ we have that $z \in L_i(y, \theta)$, as there is no profitable deviation to report $m_i = (z, \theta, \cdot)$ inducing outcome z from rule (ii). With this, $L_i(y, \theta') \cap A(\theta') \subseteq L_i(y, \theta) \cap A(\theta')$. Therefore, by strong comonotonicity, we have that $y \in F(\theta)$, a contradiction.

Case 2. Now suppose that there is a Nash equilibrium m^*, which is in rule (ii), at state θ such that $g(m^*) = y \notin F(\theta)$. It must be that $\exists j \in N$ such that, $\forall i \in N \backslash \{j\}$ we have $m_i^* = (x, \theta', n^i)$, while $m_j^* \neq (x, \theta', \cdot)$. For this to be a Nash equilibrium it must be that there is not an incentive for any agent to deviate. If $k > 1$ a deviation can lead to rule (i), (ii), or (iii), regardless, as m^* is a Nash equilibrium at θ, no agent $i \neq j$ to wish to change their report, inducing rule (iii), it must be that $y \in \operatorname{argmax}_{z \in A(\theta')} u_i(z, \theta)$. By Safe No-Veto, it must therefore be that $y \in F(\theta)$, a contradiction to $y \notin F(\theta)$. For $k = 1$ we have that a deviation can lead to rule (i), (ii), or (iv), which in the case of rule (iv) can induce any outcome. Those that can deviate to impose rule (iv) are all agents other than j. With this, we have that, as there is no incentive to deviate, that $y \in \operatorname{argmax}_{z \in \bigcup_{\theta'' \in \Theta} A(\theta'')} u_i(z, \theta)$ for all $i \in N \backslash \{j\}$. With this, it must be that $y \in \operatorname{argmax}_{z \in A(\theta')} u_i(z, \theta)$ for all $i \in N \backslash \{j\}$, and therefore by Safe No-Veto we have that $y \in F(\theta)$, a contradiction.

Case 3. Now suppose that there is a Nash equilibrium m^*, which is in rule (iii), at state θ and $g(m^*) = y \notin F(\theta)$. Suppose that $|D| < k$ and $m_i^* = (x, \theta', \cdot)$ for all agents $i \notin D$. Given this, it must be that there is no profitable deviation for any agent. As there exists a message for any player that leads to any allocation in $A(\theta')$ via rule (iii), we conclude that $y \in \operatorname{argmax}_{z\in A(\theta')} u_i(z,\theta)$ for all $i \in N$. Therefore by Safe No-Veto, we have that $y \in F(\theta)$. Now suppose that $|D| = k$. For there to be no profitable deviation, it must be that for $\forall i \in D$, $y \in \operatorname{argmax}_{z\in A(\theta')} u_i(z,\theta)$. For all agents in $i \in N\backslash D$ it must be that for any $x \in \bigcup_{\theta''\in\Theta} A(\theta'') \supseteq A(\theta')$, we have that $u_i(y,\theta) \geq u_i(x,\theta)$, as there is no profitable deviation. Given this, we conclude that $y \in \operatorname{argmax}_{z\in A(\theta')} u_i(z,\theta)$ for all $i \in N$, and therefore by Safe No-Veto we conclude that $y \in F(\theta)$, a contradiction.

Case 4. Finally, if there is a Nash equilibrium m^* at θ in rule (iv), we can see that a unilateral deviation can lead to any outcome in $\bigcup_{\theta''\in\Theta} A(\theta'')$ via rule (iv). With this, it must be that for m^* with $g(m^*) = y$ to be a Nash equilibrium in this state we have that $y \in \operatorname{argmax}_{z\in\bigcup_{\theta''\in\Theta} A(\theta'')} u_i(z,\theta)$ for all $i \in N$. Therefore, $y \in \operatorname{argmax}_{z\in A(\theta')} u_i(z,\theta)$ for some θ', and therefore by Safe No-Veto we have that $y \in F(\theta)$.

Step 3. We will now show that all Nash equilibria are safe. We consider four cases:

Case 1. If m^* is a Nash equilibrium at θ that falls into rule (i) it must be that $m_i^* = (y, \theta', n^i)$. By the previous analysis, we know that $y \in F(\theta)$. If $\theta' = \theta$, we conclude that safety is satisfied as k deviations can only lead to rule (ii) or (iii). Either way, we remain in $A(\theta)$. Now suppose that $\theta' \neq \theta$ while m^* is a Nash equilibrium at θ. Notice that regardless, k deviations must lead to remaining within $A(\theta')$ via rule (ii) or (iii). By the previous analysis, we know that this only occurs when $L_i(y,\theta') \cap A(\theta') \subseteq L_i(y,\theta) \cap A(\theta')$ for all $i \in N$. Given this, $A(\theta') \subseteq A(\theta)$ must hold for strong comonotonicity to be satisfied. Therefore any deviation from this Nash equilibrium must remain in $A(\theta') \subseteq A(\theta)$, maintaining safety.

Case 2. Now suppose that m^* is a Nash equilibrium at θ that falls into rule (ii). It must be that $\forall i \neq j$ $m_i^* = (x, \theta', n^i)$ while $m_j^* \neq (x, \theta', n^i)$. Notice that k deviations can lead to rule (i), rule (iii) if $k > 1$, and rule (iv). Notice k deviations can lead to rule (iii) for some state $\theta'' \neq \theta'$ if $k = \frac{n}{2} - 1$, depending on the report of j. Regardless, safety will require that $A(\theta) = \bigcup_{\theta''\in\Theta} A(\theta'')$ for this mechanism. To see this is implied by the condition of Safe No-Veto we only have a Nash equilibrium at such a state if $\forall i \notin N\backslash\{j\}$ they prefer $g(m^*) = y$ rather than inducing any outcome in rule (iii), in the case $k > 1$, or rule (iv), in the case, that $k = 1$. Given this, it must be that $y \in \operatorname{argmax}_{z\in A(\theta')} u_i(z,\theta)$ for all $i \notin N\backslash\{j\}$, and hence by Safe No-Veto $A(\theta) = X \Rightarrow X = \bigcup_{\theta''\in\Theta} A(\theta'')$ so safety is not violated.

Case 3. Now suppose that m^* is a Nash equilibrium at θ that falls into rule (iii), and therefore $k > 1$. It must be that all agents in $i \in N\backslash D$ for some $D \subset N$ with $|D| \leq k$, are reporting $m_i^* = (x, \theta', n^i)$. By the structure of the mechanism, k deviations can lead to rules (i), (ii) if $n = 3$ and $k = 1$ or $k \geq |D| > \frac{n}{4}$ if all those in D report $m_j = (z, \theta'', n^j)$, (iii), or (iv). With this, it is possible that for safety to be achieved we require that $A(\theta) = \bigcup_{\theta''} A(\theta'')$. Notice that for $y = g(m^*)$ to be a Nash equilibrium at state θ, by the previous analysis it must be that $y \in \operatorname{argmax}_{z \in A(\theta')} u_i(z, \theta)$ for all $i \in N$. With this, it must then be that by safety no veto $A(\theta) = \bigcup_{\theta'' \in \Theta} A(\theta'')$. Therefore Safety is necessarily achieved.

Case 4. Finally, suppose that m^* is a Nash equilibrium at θ with $g(m^*) = y$. Note that by the rules of the mechanism, k deviations can lead to any outcome via rule (iv). If we have a Nash equilibrium within this rule, it must be that $y \in \operatorname{argmax}_{z \in \bigcup_{\theta'' \in \Theta} A(\theta'')} u_i(z, \theta)$ for all $i \in N$, as else any agent could deviate to induce any outcome in $\bigcup_{\theta'' \in \Theta} A(\theta'')$ they wish via announcing a higher integer. With this, we conclude that it must be that $y \in \operatorname{argmax}_{z \in A(\theta')} u_i(z, \theta)$ for any $A(\theta')$ such that $y \in A(\theta')$. With this, by Safe No-Veto, we conclude that $A(\theta) = X \Rightarrow \bigcup_{\theta'' \in \Theta} A(\theta'') = X$, and therefore Safety is achieved. ■

Proof of Lemma 1: Take $\theta, \theta' \in \Theta$ such that $f(\theta) = x \neq f(\theta')$. Let agent i be such that $\theta_i \neq \theta_i'$. Without loss of generality, suppose that $\theta_i' > \theta_i$. We need to show $\exists y \in A(\theta)$ such that $y \in L_i(f(\theta), \theta)$ while $y \notin L_i(f(\theta), \theta')$. By Taylor's theorem, $\exists \epsilon > 0$ such that for $\mathcal{N}_\epsilon(x)$ the remainder term of the 1 Taylor expansion is sufficiently small to preserve inequalities. Therefore we need to show that there exists $y \in \mathcal{N}_\epsilon(x)$ such that $(y_1^i - x_1^i)\frac{\partial u_i(f(\theta),\theta_i)}{\partial x_1^i} + (y_2^i - x_2^i)\frac{\partial u_i(f(\theta),\theta_i)}{\partial x_2^i} < 0$ while $(y_1^i - x_1^i)\frac{\partial u_i(f(\theta),\theta_i')}{\partial x_1^i} + (y_2^i - x_2^i)\frac{\partial u_i(f(\theta),\theta_i')}{\partial x_2^i} > 0$ as $\mathcal{N}_\epsilon(f(\theta)) \subseteq A(\theta)$. With some rearranging we find $\frac{\frac{\partial u_i(f(\theta),\theta^i)}{\partial x_2^i}}{\frac{\partial u_i(f(\theta),\theta_i)}{\partial x_1^i}} < -\frac{y_1^i - x_1^i}{y_2^i - x_2^i} < \frac{\frac{\partial u_i(f(\theta),\theta_i')}{\partial x_2^i}}{\frac{\partial u_i(f(\theta),\theta_i')}{\partial x_1^i}}$, which as $\theta_i' > \theta_i$ is satisfied by single crossing, as we can find $-\frac{y_1^i - x_1^i}{y_2^i - x_2^i}$ satisfying the inequalities needed in the neighbourhood. ■

Proof of Proposition 2: Let each agent $i \in N$ announce an outcome, which excludes all reports that would be their maximal allocation, and the state. Therefore $M_i = int(X) \times \Theta$, with typical element $m_i = (x(i), \theta(i))$ Let $g(m)$ be as follows:

(i) If $m_i = (x(i), \theta(i))$ is such that $\theta(i) = \theta \quad \forall i \in N$ then $g(m) = f(\theta)$.

(ii) If $m_i = (x(i), \theta(i))$ is such that $\theta(i) = \theta \quad \forall i \in N \backslash \{j\}$ where $m_j = (x(j), \theta')$, $\theta' \neq \theta$

$$g(m) = \begin{cases} x(j) & \text{if } x(j) \in L_j(f(\theta), \theta) \cap \mathcal{N}_{\frac{\epsilon}{2}}(f(\theta)) \\ f(\theta) & \text{if } x(j) \notin L_j(f(\theta), \theta) \cap \mathcal{N}_{\frac{\epsilon}{2}}(f(\theta)) \end{cases}$$

(iii) If $\exists D \subset N$ such that $k \geq |D| > 1$, where $m_i = (x(i), \theta(i))$ and $\theta(i) = \theta, \forall i \in N \backslash D$, then $g(m)$ is constructed by the following: Let ϵ be fixed across agents such that $\mathcal{N}_\epsilon(f(\theta)) \subseteq A(\theta)$. $\forall i \in D$ let $\tilde{x}(i) = x(i)$ if $x(i) \in \mathcal{N}_{\frac{\epsilon}{|D|}}(f(\theta))$. $\tilde{x}(i) = \lambda^i x(i) + (1 - \lambda^i) f(\theta)$ such that $d(f(\theta), \tilde{x}(i)) = \frac{\epsilon}{|D|+1}$, $\lambda^i \in (0,1)$ otherwise. where Now let $g(m) = f(\theta) + \sum_{i \in D}(\tilde{x}(i) - f(\theta))$.

(iv) Otherwise, let $g(m) = \frac{1}{n}\sum_{i \in N} x(i)$.

Step 1. First to show that $x = f(\theta)$ is a Nash Equilibrium at θ. Consider m^* satisfying rule (i) Any unilateral deviation of agent i leads to rule (ii), where the only way to change the allocation is in $L_i(f(\theta), \theta)$, which cannot give a strictly higher utility by definition. Therefore all m^* satisfying rule (i) are Equilibria.

Step 2. We want to show that $\nexists m^*$ that is an Equilibrium at θ with $g(m^*) \neq f(\theta)$.

Case 1: Suppose that there is an Equilibrium in Rule (i) where $g(m^*) \neq f(\theta)$, where the true state is θ. It follows that all agents are announcing some state $\theta' \neq \theta$. With this, there exists some agent who announces their own type to be $\theta_j(j) = \theta'_j \neq \theta_j$. For this agent $\exists x_j$ s.t. $x_j \in L_j(f(\theta'), \theta') \cap \mathcal{N}_{\frac{\epsilon}{2}}(f(\theta'))$ while $x_j \notin L_j(f(\theta'), \theta) \cap \mathcal{N}_{\frac{\epsilon}{2}}(f(\theta'))$ by the same logic as lemma 1 via the single crossing condition. Therefore m^* cannot be a Nash Equilibrium.

Case 2: There are no Nash Equilibria for any θ in rule (ii). Suppose that m^* is an equilibrium at that θ where for all $i \in N \backslash \{j\}$ we have that $m_i = (x(i), \theta(i))$ with $\theta(i) = \theta'$ while $m_j = (x(j), \theta(j))$ with $\theta(j) \neq \theta'$. Regardless of whether $g(m^*) = f(\theta)$ or $g(m^*) = x(j)$, notice that any agent $i \neq j$ can induce an increase in both dimensions of the bundle by announcing $m_i = (x'(i), \theta'(i))$, where $\theta'(i) \neq \theta'$ and $x'(i)$ such that $x'_j(i) = f_j(\theta)$ and $x'_i(i)$ is chosen such that $x'(i) \in \mathcal{N}_{\frac{\epsilon}{2}}(f(\theta))$ and $\frac{x_i^{k,\prime}(i) + \tilde{x}_j^k(j)}{2} > f_i^k(\theta)$, which is achievable by the construction of rule (iii). As u_i is strictly increasing, m^* is not a Nash equilibrium.

Case 3: There cannot be an Equilibrium in Rule (iii), any agent $i \in D$ can announce an allocation to the north east of $\tilde{x}(i)$ such that $x(i) \in \mathcal{N}_{\frac{\epsilon}{|D|}}(f(\theta))$, leading to rule (iii) or (iv), regardless, monotonically increase their allocation.

Case 4: The final case is within rule (iv). Again, this cannot be an Equilibrium, as agents can deviate to announcing an allocation to the north east of the current one, leading to rule (iv). This deviation is profitable given the assumption of increasing utility. As the message can only be interior in X, such a profitable deviation always exists.

Step 3: Notice all Equilibria lie in Rule (i). Further, any such equilibrium m^* at θ lead to $g(m^*) = f(\theta)$ by Case 1 of Step 2. k deviations that remains in rule (i) must lead to the same allocation, and therefore safety is guaranteed. k deviations that lead to rule (ii) lead to allocations in $\mathcal{N}_{\frac{\epsilon}{2}}(f(\theta)) \subset \mathcal{N}_\epsilon(f(\theta)) \subseteq A(\theta)$ and therefore safety is maintained. The only check needed for this is that rule (iii) lies within an ϵ neighbourhood of $f(\theta)$, and therefore within $A(\theta)$. To see this, notice that:

$$d(f(\theta), g(m)) = d\left(f(\theta), f(\theta) + \sum_{i\in D}(\tilde{x}(i) - f(\theta))\right) = \|\sum_{i\in D}(\tilde{x}(i) - f(\theta))\|$$
$$\leq \sum_{i\in D}\|\tilde{x}(i) - f(\theta)\| = \sum_{i\in D} d(f(\theta), \tilde{x}(i)) < |D|\frac{1}{|D|}\epsilon = \epsilon$$

(the weak inequality comes from the triangle inequality). Hence, $g(m) \in \mathcal{N}_\epsilon(f(\theta))$ for any m within rule (iii) that is k deviations from an equilibrium at θ. ■

Proof of Proposition 4: If $|X| \leq n$, by richness $\exists \theta \in \Theta$ such that for every $x \in X$ $\exists i \in N$ such that $\{x\} = \text{argmin}_{y\in X} u_i(y, \theta)$. Hence if A is minimally safeguarding then $X^*(\theta) = \emptyset$ and therefore no SCC can be safely $\mathcal{C}$-implemented for any $k \geq 1$ and any $\mathcal{C}$.■

Proof of Proposition 5: If it is not the case that $A(\theta) = X$ for some θ, then it must be that some $x \in X$ is not in $A(\theta)$. By surjectivity, there is some state where $x = f(\theta')$, and $x \neq z = f(\theta)$. By richness, $\exists \theta'' \in \Theta$ where x is the top ranked alternative for all players, while z is second ranked for all players. Hence, by Comonotonicity, both z and x are chosen by the SCF at θ''. But since $x \neq z$, and we have a SCF, this is a contradiction. ■

Proof of Proposition 3: Let $X = N \cup \{0\}$, where 0 represents the good not being allocated. For each $\theta \in \Theta$ let $\theta \in \mathbb{R}^n_+$ denote the vector of agents' values. Let $M_i = X \times \mathbb{R}^n_+$ for all $i \in N$ with a typical message $m_i = (j, v) \in N \cup \{0\} \times \mathbb{R}^n_+$. Let $g(m)$ be as follows:

(i) If $\forall i \in N$ $m_i = (j', v)$ with $v = \theta \in \Theta$ and $j' = f(\theta)$ then $g(m) = j' = f(\theta)$.

(ii) If $m_i = (j', v)$ $\forall i \in N \backslash \{j\}$ with $v = \theta \in \Theta$ and $f(\theta) = i'$ and $m_j = (l, \cdot)$, then

$$g(m) = \begin{cases} l & \text{if } l \in [L_j(j', \theta) \cap \tilde{A}(\theta)] \backslash \{j'\} \\ \emptyset & \text{if } l \notin [L_j(j', \theta) \cap \tilde{A}(\theta)] \backslash \{j'\} \end{cases}$$

(iii) If $m_i = (j', v)$ such that $v = \theta \in \Theta$ and $j' = f(\theta)$ for $\forall i \in N \backslash D$, $2 \leq |D| < \frac{n}{2}$ such that $\forall j \in D$ $m_j = (l^j, \cdot)$, $l^j \neq j'$ then

$$g(m) = \begin{cases} l^{i^*} & \text{if } D^*(\theta, D) \neq \emptyset \\ j' & \text{if } D^*(\theta, D) = \emptyset \end{cases}$$

where $D^*(\theta, D) = \{j \in D | l^j \in \tilde{A}(\theta)\}$ and $i^* = \min\{i \in D^*(\theta, D) | v_i^i \geq v_j^j \quad j \in D^*(\theta, D)\}$.

(iv) otherwise let $g(m) = l^{i^*}$ where $m_i = (l^i, \cdot)$ and $i^* = \min\{i \in N | v_i^i \geq v_j^j \quad j \in N\}$.

Where $\tilde{A}(\theta) = \bigcap_{\theta' \in \Theta | f(\theta') = f(\theta)} A(\theta')$.

Notice that, at state θ, with messages that fall into rule (i) with $m^* = (j', \theta)$, m^* is a Nash equilibrium, since any deviation from m^* either leads to the good not being allocated or it must be that a less deserving agent receives the good. To show all Nash Equilibria are safe, we will do so by showing that rule $i)$ constitute the only Nash Equilibria, and always allocate the $f(\theta)$ at state θ.

Suppose that there is a Nash Equilibrium in rule $ii)$ m^* at state θ. Let $m_i^* = (j', \theta')$ for all $i \neq j$ and $m_j^* = (l, \cdot)$. It must be either $g(m^*) = l \in \tilde{A}(\theta')$, $l \in N \backslash \{j'\}$, or $g(m^*) = 0$. Suppose that $j = j'$. Here there is a profitable deviation to announce $m_j = (j', \theta')$ and be allocated the good, which cannot be case under rule (ii). Suppose instead that $j \neq j'$. Let $i = j'$, who can announce $m_i = (i, v'')$ such that v_i'' is strictly higher than the i^{th} (or equivalently $j'^{,th}$) component of θ' and receive the good by inducing rule (iii).

As all agents prefer to have the good allocated to themselves, there can be no Equilibria in rule $iii)$ and $iv)$. To see that in the case of rule (iii) there is no Nash equilibrium, suppose that the message of $|N| - k$ agents is $m_i = (j', v')$, with $v' = \theta'$ and $f(\theta') = j'$, while m^* is a Nash equilibrium. Given that there is some agent $j \in \tilde{A}(\theta')$ such that $g(m^*) \neq j$ by (A.3.). Such an agent prefers to have the good allocated to themself, they can announce $m_j = (j, v'')$, such that $v_j'' = \max_{i \neq j} v_i^i + \epsilon$, and therefore would induce that the good is allocated to them. For rule (iv), but any agent who is not allocated the good could deviate.

Suppose that there is some Nash Equilibrium in rule i) m^* at θ such that, for some θ' we have $g(m^*) = f(\theta') = j' \neq f(\theta)$. j' is undeserving. Any agent can announce $l = 0$ (or any $l \notin A(\theta)$), which given rule (ii) and (P.2.), induces no agent to receive the good, as is not preferred at θ'. However, this is preferred at θ as reverting to the empty allocation is attainable and by assumption gives a higher payoff than an undeserving agent.

Notice that they all lie within rule (i) with $m_i^* = (j', \theta)$ at state θ', where j' has the highest valuation in state θ'. Up to k deviations can only lead to rules (ii) or (iii), where the majority still announces (j', θ). With this, we remain in $\tilde{A}(\theta) \subseteq A(\theta')$. ■

APPENDIX B: ON THE GAP BETWEEN WEAK AND STRONG COMONOTONICITY

Strong and *Weak* Comonotonicity coincide for SCFs, but when the SCC is not single valued, there is a gap between necessary and sufficient conditions. In this appendix we show that a stronger condition than Weak Comonotonicity is necessary and almost sufficient, thereby reducing the gap between necessity and sufficiency. Similar to Moore and Repullo (1990)'s 'Condition μ', this condition relies on identifying which sub-correspondences of A are used, within an implementing mechanism, to support each of the different allocations in the SCC. Like Moore and Repullo (1990)'s 'Condition μ' compared Maskin Monotonicity, however, this condition too is harder to check than Weak Comonotonicity.

Specifically, let $\mathcal{M} = \langle (M_i)_{i\in N}, g\rangle$ be a mechanism that (A,k)-Safe Implements F. For any θ and $x \in F(\theta)$, let $NE(x,\theta) \subseteq M$ denote the (non-empty) set Nash equilibria at state θ that induce x. Then, for each $m^*(x,\theta) \in NE(x,\theta)$ we know that (i) $x = g(m^*)$, and (ii) $g(m) \in A(\theta)$ for any $m \in B_k(m^*)$ (i.e., for any m that is within k deviations from m^*). Next, let $G^k(x,\theta) := \cup_{m^*\in NE(x,\theta)} B_k(m^*)$. By definition of Safety, $G^k(x,\theta) \subseteq A(\theta)$. Essentially, for each θ and $x \in F(\theta)$, $G^k(x,\theta)$ is the subset of $A(\theta)$ that consists of all the allocations that are used to 'sustain' the implementation of outcome x.

Notice that, for $k = 1$, the set $G^1(x,\theta)$ consists of the set of allocations that can be induced by *unilateral* deviations from one of the Nash equilibria $m^* \in NE(x,\theta)$, and similar to Moore and Repullo (1990), let $C_i(x,\theta) \subseteq G^1(x,\theta)$ denote the set of allocations that can be induced by unilateral deviations of player i alone. Then, $C_i(x,\theta) \subseteq G^k(x,\theta) \subseteq A(\theta)$ and $x \in \text{argmax}_{y\in C_i(x,\theta)} u_i(y,\theta)$ for all $i \in N$.[20] Next notice that if for some θ' it holds that

[20]To see why the latter condition holds, for any $m^* \in NE(x,\theta)$, let $C_i(m^*) := \{y \in X : \exists m_i \in M_i \text{ such that } y = g(m_i, m^*_{-i})\}$. Then, $C_i(x,\theta) = \cup_{m^*\in NE(x,\theta)} C_i(m^*)$, and since $x \in \text{argmax}_{y\in C_i(m^*)} u_i(y,\theta)$ for all i and for all $m^* \in NE(x,\theta)$, it follows that $x \in \text{argmax}_{y\in C_i(x,\theta)} u_i(y,\theta)$ for all $i \in N$.

$x \in \mathrm{argmax}_{y \in C_i(x,\theta)} u_i(y, \theta')$ for all i, then all $m^* \in NE(x,\theta)$ are also equilibria at θ', and hence $NE(x,\theta) \subseteq NE(x,\theta')$. It follows that (i) $x \in F(\theta')$, and (ii) $G^k(x,\theta) \subseteq G^k(x,\theta')$.[21] With this, we obtain that the following condition is necessary:

DEFINITION 13: (A, F) *satisfy the* Safe-μ Condition *if there exist correspondences* $G : X \times \Theta \rightrightarrows X$ *and* $C_i : X \times \Theta \rightrightarrows X$ *such that* $G(x,\theta) \subseteq A(\theta)$ *and* $C_i(x,\theta) \subseteq L_i(x,\theta) \cap G(x,\theta)$ *for all* i, θ *and* $x \in F(\theta)$*, which satisfy the following: if* $\theta, \theta' \in \Theta$ *and* $x \in F(\theta)$ *are such that* $C_i(x,\theta) \subseteq L_i(x,\theta')$ *for all* i*, then: (i)* $x \in F(\theta')$*, and (ii)* $G(x,\theta) \subseteq G(x,\theta')$*.*

THEOREM 4: F *is* (A,k)*-Safe Implementable* only if *the Safe-*μ *Condition is satisfied. If, moreover,* A *is* maximally safe*, then* $\cup_{x \in F(\theta)} G(x,\theta) = A(\theta)$ *for each* θ*.*

The gap between Comonotonicity and Def. 13 is analogous to the gap between Monotonicity and Condition μ of Moore and Repullo (1990). Similarly, under the appropriate No-Veto condition, the Safe-μ Condition can be shown to be sufficient for (A,k)-Safe Implementation when $k < \frac{n}{2}$. All the results in Section 4.1 would also hold under the suitable adaptations of No Unanimity and No Total Indifference, and hence a tight characterization can be provided for general SCC in those environments.

This condition also identifies the exact source of the gap between strong and weak Comonotonicity when the SCC is non-single valued: if, for some state θ, $F(\theta)$ contains multiple allocations, say $x, x' \in F(\theta)$, different subsets of $A(\theta)$ may be used to sustain them, namely $G^k(x,\theta)$ and $G^k(x',\theta)$. When x 'climbs up' from θ to θ', then it must be that the $x \in F(\theta')$ and that all $G^k(x,\theta)$ must also be acceptable at θ'. However, unless this happens *for all* allocations in $F(\theta)$ (cf. point 2 in Def. 5), we cannot conclude that $A(\theta) \subseteq A(\theta')$, even under maximal (A,k)-Safe Implementation. We may only conclude that some subset of allocations of $A(\theta)$ are a subset of $A(\theta')$ (more precisely, that $G^k(x,\theta) \subseteq G^k(x,\theta') \subseteq A(\theta')$). Clearly, $A(\theta) \subseteq A(\theta')$ would follow immediately if $G^k(x,\theta) = A(\theta)$ for all $\theta \in \Theta$ and $x \in F(\theta)$, in which case in fact Safe-μ boils down precisely to Strong Comonotonicity. But when the G^k are strict subcorrespondences of A, then the condition becomes much harder to check. For these reasons, we elect to provide Weak and Strong Comonotonicity as more transparent and easy to check conditions.

---

[21] Point (i) follows from implementation; point (ii) from the fact that $NE(x,\theta) \subseteq NE(x,\theta')$.

APPENDIX C: ON THE RELATIONSHIP BETWEEN SAFE IMPLEMENTATION AND FAULT TOLERANT IMPLEMENTATION

In this appendix we provides two examples to show that, despite their similar motivation, Safe Implementation and Fault Tolerant Implementation of Eliaz (2002) are distinct and non-nested notions. We first recall the definition of Fault Tolerant Nash Equilibrium:

DEFINITION 14: *A k-Fault Tolerant Nash Equilibrium (k-FTNE) for the instance (θ, k) is a profile of messages $m^* \in M$ having the property that $\forall i \in N$, $\forall m_i \in M_i$, $\forall m_D \in M_D$ and $\forall D \subseteq N$ such that $|D| \leq k$:*

$$u_i(g(m_i^*, m^*_{N\setminus\{D\cup\{i\}\}}, m_D), \theta) \geq u_i(g(m_i, m^*_{N\setminus\{D\cup\{i\}\}}, m_D), \theta).$$

Let $\mathcal{C}_k^{\mathcal{M}}(\theta)$ denote the set of k-FTNEin mechanism $\mathcal{M}$ at state θ.

The definition of k-Fault Tolerant Implementation (k-FTI) requires that the set of k-Fault Tolerant Implementation coincide with the designer's desired outcomes, as dictated by the social choice correspondence (SCC), and additionally that the set of outcomes that are reachable within k deviations from any such equilibria are also within the SCC.

DEFINITION 15: *Let $\langle N, \Theta, X, (u_i)_{i\in N}\rangle$ be an environment. The SCC $F : \Theta \to 2^X\setminus\{\emptyset\}$ is k-Fault Tolerant implemented by $g : M \to X$, if $\forall\theta \in \Theta$, $\forall\ m^* \in \mathcal{C}_k^{\mathcal{M}}(\theta)$: (i) $g(\mathcal{C}_k^{\mathcal{M}}(\theta)) = F(\theta)$; and (ii) $g(B(m^*, k)) \subseteq F(\theta)$.*

Eliaz (2002) introduced two key conditions, *k-monotonicity* and *weak k-monotonicity*, and shows that the first is necessary for k-FTI in the case of SCF, and the second for SCC. (In the case of SCFs, the two notions coincide):

DEFINITION 16: *A SCC $F : \Theta \to 2^X\setminus\{\emptyset\}$ is k-**monotonic** if, whenever $x \in F(\theta)$ and $x \notin F(\theta')$, there exists $D \subset N$ and $\exists y \in X$ such that $|D| \geq k+1$, every $i \in M$ satisfies $u_i(x, \theta) \geq u_i(y, \theta_i)$ and at least one player $j \in M$ satisfies $u_j(y, \theta'_j) > u_j(x, \theta'_j)$.*

DEFINITION 17: *A SCC $F : \Theta \to 2^X\setminus\{\emptyset\}$ is **weakly** k-**monotonic** if, whenever $F(\theta) \not\subseteq F(\theta')$, there exists $D \subseteq N$ have at least $k+1$ players and $\exists y \in X$ such that, for every player $i \in D$, there is an outcome $x^i \in F(\theta)$ satisfying $u_i(x^i, \theta) \geq u_i(y, \theta)$, and for at least one of these players $j \in D$, $u_j(y, \theta') > u_j(x^j, \theta')$*

The next example shows that a non-(Maskin) monotonic SCC may be 1-FTI. This illustrates three things, all of which were discussed in the main text: first, 1-FTI is possible when Safe Implementation is not, regardless of the acceptability correspondnece; second, since 0-FTI coincides with Nash Implementation, that k-FTI need not imply $(k-1)$-FTI; third, k-FTI cannot be seen as an extra desideratum on top of Nash Implementation.

EXAMPLE 6—A rule that is Implementable in 1-Fault Tolerant Equilibrium but not Nash Equilibrium: *Take* $N=\{1,2,3\}$, $\Theta=\{\theta_1,\theta_2\}$ $X=\{a,b,c,d,e\}$, $u_i(x,\theta_2)=0, \forall x\in X, \forall i\in N$, *and let utilities* $u(x,\theta_1)=(u_1(x,\theta_1),u_2(x,\theta_1),u_3(x,\theta_1))$ *of each outcome at state* θ_1*be as follows:* $u(a,\theta_1)=(1,1,1)$, $u(b,\theta_1)=(1,0,1)$, $u(c,\theta_1)=(0,1,1)$,$u(d,\theta_1)=(0,0,0)$, *and* $u(e,\theta_1)=(1,1,2)$. *Finally, the SCC is* $F(\theta_1)=\{a,b,c\}$ *and* $F(\theta_2)=X$.

Note this SCC violates (Maskin) monotonicity: since $X=L_i(e,\theta_1)=L_i(e,\theta_2)$ *for all* i, *monotonicity would require* $e\in F(\theta_1)$. *Hence, this rule is not Nash Implementable, and thus not Safe Implementable, for any acceptability correspondence or* k. *Yet, the following mechanism achieves* 1*-FTI of this SCC: For each* i, $M_i=\{1,2,3\}$, *and* $g(m)$ *is as follows:*

$m_3=1$

		m_2 = 1	2	3
m_1	1	a	c	c
	2	b	d	d
	3	b	d	d

$m_3=2$

		m_2 = 1	2	3
m_1	1	c	c	c
	2	b	d	d
	3	b	d	d

$m_3=3$

		m_2 = 1	2	3
m_1	1	b	c	c
	2	b	d	d
	3	b	d	e

TABLE C.I

A 1-FT IMPLEMENTING MECHANISM: 1 CHOOSES THE ROW MESSAGE, 2 CHOOSES THE COLUMN MESSAGE, AND 3 CHOOSES THE TABLE MESSAGE; THE OUTCOME $g(m)$ INDUCED BY EACH MESSAGE PROFILE IS REPRESENTED IN THE CORRESPONDING CELL.

At state θ_1, *this mechanism induces the following game:*

		m_2		
		1	2	3
m_1	1	(1,1,1)	(1,0,1)	(1,0,1)
	2	(0,1,1)	(0,0,0)	(0,0,0)
	3	(0,1,1)	(0,0,0)	(0,0,0)

$m_3 = 1$

		m_2		
		1	2	3
m_1	1	(1,0,1)	(1,0,1)	(1,0,1)
	2	(0,1,1)	(0,0,0)	(0,0,0)
	3	(0,1,1)	(0,0,0)	(0,0,0)

$m_3 = 2$

		m_2		
		1	2	3
m_1	1	(0,1,1)	(1,0,1)	(1,0,1)
	2	(0,1,1)	(0,0,0)	(0,0,0)
	3	(0,1,1)	(0,0,0)	(1,1,2)

$m_3 = 3$

TABLE C.II

THE INDUCED GAME AT STATE θ_1.

First note that $m = (1,1,1)$ is a 1-FTNE that induces a: under any unilateral deviations of some of i's opponents, message $m_i = 1$ still yields a payoff at least as high as that obtained from sending a different message, while at the same time ensuring outcomes consistent with the SCC at that state (namely, b or c).

Second, $m = (1,1,2)$ induces c and is also a 1-FTNE: if any one opponent deviates, no player can increase their utility by also deviating, and any unilateral deviation still results in outcomes (a or c) consistent with the SCC at θ_1.

The same is true of $m = (1,1,3)$ which induces b. Further, it can be seen that there are no other 1-FTNE in this game. Hence, each of the outcomes in $F(\theta_1)$ is induced as a 1-FTNE outcome, and unilateral deviations from any such equilibrium result in outcomes within $F(\theta)$. Since implementation at state θ_2 is trivial, it follows that this mechanism 1-*FT-Implements the SCC.* $\square$

We now turn to showing there are cases where Safe Implementation is possible, even under the most restrictive case of perfect safety, while 1-FTI is not. To do so, we will show that both 1-monotonicity and weak 1-monotonicity are violated.

EXAMPLE 7: *Let there be four players $N = \{1,2,3,4\}$, three alterantives $X = \{a,b,c\}$ and two states of the world, L and R, with the SCC such that $F(L) = X$ while $F(R) = \{b,c\}$. Then, consider* perfect safety*, i.e., $A(\theta) = F(\theta)$ for all θ (see Fig. C.1):*

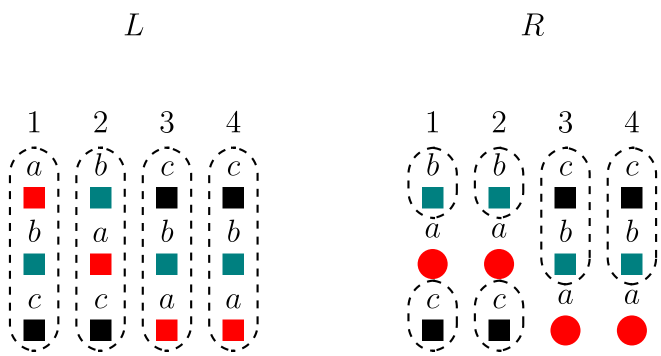

FIGURE C.1.—Let $F(L) = X = A(L)$ and $F(R) = \{b,c\} = A(R)$. The preferences are represented top to bottom. For instance, in state L player 1 has the ordering $a \succ b \succ c$.

First notice that comonotonicity *holds. To see this, we need to consider that $a \in F(L)$ but $a \notin F(R)$. But since $L_1(a,R) \cap A(R) = X$ while $L_1(a,L) \cap A(R) = \{a,c\}$, we have $L_i(a,R) \cap A(R) \nsubseteq L_i(a,L) \cap A(R)$ for some i, and hence comonotonicity does not require that $a \in F(R)$. Further, as Safe No-Veto is not violated F is $(A,1)$-Safe implementable with $A(\theta) = F(\theta)$ for all θ.*

1-monotonicity, however, does not hold. For it to hold, it must be that two players at state R prefer some other common allocation to a, and one such agent reverses their preferences at state L. But a is worst ranked for 3 and 4 in both L and R, and hence the only possible candidate is agent 2, who only prefers a to c in L. As neither 1 or 2 have a preference reversals around a and c from L to R, 1-monotonicity does not hold. Since 2's preferences do not change, the same logic also applies to show that weak 1-monotonicity does not hold either, as there is no preference reversal around the only commonly dominated outcome c in any of the outcomes in $F(L)$ for 1 and 2. □

REFERENCES

Arya, A., Glover, J., and Rajan, U. (2000). Implementation in principal–agent models of adverse selection. *Journal of Economic Theory*, 93(1):87–109. [11]

Barlo, M. and Dalkıran, N. A. (2021). Implementation with missing data. *Working Paper, Sabanci University*. [28]

Barlo, M. and Dalkıran, N. A. (2023). Behavioral implementation under incomplete information. *Journal of Economic Theory*, 213:105738. [28]

Ben-Porath, E., Dekel, E., and Lipman, B. L. (2019). Mechanisms with evidence: Commitment and robustness. *Econometrica*, 87(2):529–566. [28]

Benoît, J.-P. and Ok, E. A. (2008). Nash implementation without no-veto power. *Games and Economic Behavior*, 64(1):51–67. [21]

Bergemann, D. and Morris, S. (2005). Robust mechanism design. *Econometrica*, 73(6):1771–1813. [28]

Bergemann, D. and Morris, S. (2009a). Robust implementation in direct mechanisms. *The Review of Economic Studies*, 76(4):1175–1204. [28]

Bergemann, D. and Morris, S. (2009b). Robust virtual implementation. *Theoretical Economics*, 4(1). [28]

Bergemann, D., Morris, S., and Tercieux, O. (2011). Rationalizable implementation. *Journal of Economic Theory*, 146(3):1253–1274. [2]

Bochet, O. (2007). Nash implementation with lottery mechanisms. *Social Choice and Welfare*, 28(1):111–125. [21]

Bochet, O. and Maniquet, F. (2010). Virtual nash implementation with admissible support. *Journal of mathematical economics*, 46(1):99–108. [22, 27]

Bochet, O. and Tumennasan, N. (2023a). Defaults and benchmarks in mechanism design. *Working Paper*. [2, 28]

Bochet, O. and Tumennasan, N. (2023b). Resilient mechanisms. *Working Paper*. [2, 28]

Crawford, V. P. (2021). Efficient mechanisms for level-k bilateral trading. *Games and Economic Behavior*, 127:80–101. [6, 28, 29]

De Clippel, G. (2014). Behavioral implementation. *American Economic Review*, 104(10):2975–3002. [2, 6, 28, 29]

De Clippel, G., Saran, R., and Serrano, R. (2019). Level-k mechanism design. *The Review of Economic Studies*, 86(3):1207–1227. [2, 6, 28, 29]

Dutta, B. and Sen, A. (2012). Nash implementation with partially honest individuals. *Games and Economic Behavior*, 74(1):154–169. [20]

Eliaz, K. (2002). Fault tolerant implementation. *The Review of Economic Studies*, 69(3):589–610. [2, 4, 6, 23, 24, 25, 26, 27, 29, 30, 41]

Gavan, M. J. and Penta, A. (2024). Safe implementation. *BSE working paper series*, 1363. [19, 21, 24, 26]

Hayashi, T. and Lombardi, M. (2019). Constrained implementation. *Journal of Economic Theory*, 183:546–567. [27]

Hong, L. (1995). Nash implementation in production economies. *Economic Theory*, 5(3):401–417. [27, 28, 30]

Hong, L. (1998). Feasible bayesian implementation with state dependent feasible sets. *Journal of Economic Theory*, 80(2):201–221. [28, 30]

Jackson, M. O. (1991). Bayesian implementation. *Econometrica: Journal of the Econometric Society*, pages 461–477. [30]

Jackson, M. O. (1992). Implementation in undominated strategies: A look at bounded mechanisms. *The Review of Economic Studies*, 59(4):757–775. [30]

Jackson, M. O. and Palfrey, T. R. (2001). Voluntary implementation. *Journal of Economic Theory*, 98(1):1–25. [27]

Jain, R., Lombardi, M., and Penta, A. (2024). Strategically robust implementation. *Working Paper*. [2, 26]

Kartik, N. and Tercieux, O. (2012). Implementation with evidence. *Theoretical Economics*, 7(2):323–355. [11, 18, 19, 28]

Kartik, N., Tercieux, O., and Holden, R. (2014). Simple mechanisms and preferences for honesty. *Games and Economic Behavior*, 83:284–290. [20]

Kneeland, T. (2022). Mechanism design with level-k types: Theory and an application to bilateral trade. *Journal of Economic Theory*, 201:105421. [28]

Kunimoto, T., Saran, R., and Serrano, R. (2024). Interim rationalizable implementation of functions. *Mathematics of Operations Research, forthcoming*, 49:1791–1824. [2]

Kunimoto, T. and Serrano, R. (2019). Rationalizable implementation of correspondences. *Mathematics of Operations Research*, 44:1326–1344. [2]

Lombardi, M. and Yoshihara, N. (2020). Partially-honest nash implementation: a full characterization. *Economic Theory*, 70(3):871–904. [20]

Maskin, E. (1977). Nash equilibrium and welfare optimality. *mimeo, M.I.T.* [2]

Maskin, E. (1999). Nash equilibrium and welfare optimality. *Review of Economic Studies*, 66(1):23–38. [2, 7, 11, 30]

Maskin, E. and Sjöström, T. (2002). Implementation theory. *Handbook of social Choice and Welfare*, 1:237–288. [2]

Matsushima, H. (2008). Role of honesty in full implementation. *Journal of Economic Theory*, 139(1):353–359. [20]

Mirrlees, J. A. (1976). Optimal tax theory: A synthesis. *Journal of public Economics*, 6(4):327–358. [11]

Moore, J. and Repullo, R. (1988). Subgame perfect implementation. *Econometrica: Journal of the Econometric Society*, pages 1191–1220. [2]

Moore, J. and Repullo, R. (1990). Nash implementation: A full characterization. *Econometrica*, 58(5):1083–1099. [21, 39, 40]

Muller, E. and Satterthwaite, M. A. (1977). The equivalence of strong positive association and strategy-proofness. *Journal of Economic Theory*, 14(2):412–418. [25]

Ollár, M. and Penta, A. (2017). Full implementation and belief restrictions. *American Economic Review*, 107(8):2243–77. [28, 30]

Ollár, M. and Penta, A. (2022). Efficient full implementation via transfers: Uniqueness and sensitivity in symmetric environments. In *AEA Papers and Proceedings*, volume 112, pages 438–43. [28, 30]

Ollár, M. and Penta, A. (2023). A network solution to robust implementation: The case of identical but unknown distributions. *Review of Economic Studies*. [28, 30]

Postlewaite, A. and Wettstein, D. (1989). Feasible and continuous implementation. *The Review of Economic Studies*, 56(4):603–611. [3, 9, 27, 28, 30]

Renou, L. and Schlag, K. (2011). Implementation in minimax regret equilibrium. *Games and Economic Behavior*, 71(2):527–533. [6, 28, 29]

Shoukry, G. (2014). Safety in mechanism design and implementation theory. *Available at SSRN 2478655*. [27]

Shoukry, G. F. (2019). Outcome-robust mechanisms for nash implementation. *Social Choice and Welfare*, 52(3):497–526. [2, 8, 27]

Spence, A. M. (1980). Multi-product quantity-dependent prices and profitability constraints. *The Review of Economic Studies*, 47(5):821–841. [11]

Tumennasan, N. (2013). To err is human: Implementation in quantal response equilibria. *Games and Economic Behavior*, 77(1):138–152. [6, 28, 29]